Data Science
CA3 (Mini Project)

Submitted by

Asita Ganatra, 22070521003

B. Tech Computer Science and Engineering

Under the Guidance of

Dr. Piyush Chauhan

॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR

Wathoda, Nagpur
2025

This certifies that the DS EDA report, "Analysis and Forecasting of Credit by Scheduled Commercial Banks in India (2010-2023)," submitted by Asita Ganatra, 22070521003, partially satisfies the requirements for the DS Project and is an authentic record of work completed under my supervision. This project's contents, in whole or in part, have not been lifted from another source or submitted to another university or institute for the award of a degree or diploma, and they are certified.

Dr. Piyush Chauhan
DS Project Coordinator

The Report is satisfactory/unsatisfactory

Approved by

Prof. (Dr.) Nitin Rakesh
Director, SIT Nagpur

ABSTRACT

This project offers a thorough data science application for scheduled commercial banks in India to analyze credit between 2010 and 2023. The need for an integrated tool to investigate, evaluate, and forecast important banking metrics is the main issue addressed. The process entails using Streamlit to create an interactive web application that combines several machine learning paradigms. This includes (a) visualizing trends through exploratory data analysis (EDA); (b) forecasting loan amounts and account numbers through predictive modeling using Linear Regression and Random Forest models; (c) classifying bank groups and regions; (d) using KMeans, DBSCAN, and Agglomerative algorithms for unsupervised clustering to find hidden patterns; and (e) deep learning models, such as Autoencoders for dimensionality reduction and Deep Neural Networks (DNNs) for advanced classification. Key findings show that the DNN classifier achieves high accuracy (~90%) and Random Forest models (R^2 ~0.92) perform better in prediction tasks than Linear Regression (R^2 ~ 0.85). This work is important because it gives banking professionals a single, strong, and easy-to-use tool to extract useful information from complicated credit data.

TABLE OF CONTENTS

1. Keywords

Data Science, Machine Learning, Bank Credit Analysis, Predictive Modeling, Classification, Clustering, Deep Learning, Streamlit, Python

2. Introduction

• Background/Context: The Indian banking industry produces a significant amount of data on the distribution of credit, which is a crucial sign of the state of the economy. Financial institutions and policymakers must analyze this data, which covers a variety of geographic areas, bank types, and demographic groups.

• The goal of this study is to close the gap between unprocessed banking data and useful insights. We can find patterns, develop predictive models, and produce a tool that makes this complicated data understandable and practical by utilizing a wide range of data science approaches, from EDA to deep learning.

• The problem statement is to create and put into place a comprehensive data analysis system that offers predictive modeling, clustering, classification, and an interactive dashboard for Indian commercial bank credit data spanning the years 2010–2023.

• Objectives of the Project:

- Develop an interactive dashboard for banking data visualization.
- Implement predictive models for loan amount and number of accounts.
- Create classification systems to identify bank groups and regions.
- Implement clustering algorithms for customer and loan pattern discovery.
- Integrate deep learning capabilities for complex pattern recognition.

• Originality of your work: This project's uniqueness is found in its modular, all-in-one design. It is a comprehensive web-based analytical suite (developed with Streamlit) that enables a user to perform EDA, regression, classification, clustering, and deep learning on the same dataset using a single, user-friendly interface.

3. Literature Review / Related Work

In order to determine the primary predictors of bank risk, credit quality, and profitability, prior banking and finance research has made extensive use of empirical and statistical models. These studies frequently distinguish between external (macroeconomic) and internal (bank-specific) factors.
Finding the predictors of bank risk across several dimensions, such as credit, liquidity, and systemic risk, is one important field of research, as Suzdaltseva (2025) demonstrates. In order to identify reliable drivers, this study uses panel data from multiple banks. Typical internal

profitability (such as Return on Assets), capital adequacy (CAR), and operational efficiency are among the predictors that have been studied; macroeconomic factors such as GDP growth and inflation are examples of external predictors. Stronger internal metrics, like increased capital and profitability, are generally found to consistently lower bank risk.

The study by Dhakal (2024) on Nepali commercial banks is one example of a related branch of work that examines credit management practices in particular geographic areas. Regression analysis is used in this study to ascertain how credit management factors, such as the Non-Performing Loan Ratio (NPL) and Loan Loss Provision (LLP), affect bank performance, which is frequently gauged by return on equity (ROE).

Gaps with the earlier work: Although these studies offer a solid basis for analyzing bank data using statistical models, many (like Dhakal) concentrate more on establishing regression-based relationships than on machine learning's capacity for forecasting. Moreover, these studies are regionally specific (Europe, Nepal). By using a specialized data science and machine learning approach to forecast commercial bank credit data specifically within the Indian context, this project closes a gap.

4. Methodology / Proposed System

The suggested system is an interactive web application developed with the Streamlit framework in Python. Because of the system's modular architecture, users can switch between various analytical tasks.

Algorithms and Formulations:

- **Predictive Modeling (Regression):**
  - **Linear Regression:** Used as a baseline model to predict log-transformed amount_outstanding and no_of_accounts.
  - **Random Forest Regressor:** An ensemble model used for its high accuracy and ability to capture non-linear relationships.
- **Classification Analysis:**
  - **Logistic Regression:** Used as a linear baseline for classifying bank_group and region.
  - **Random Forest Classifier:** An ensemble model for robust, high-accuracy classification.
- **Clustering Analysis:**
  - **KMeans:** For partitioning data into k distinct clusters.
  - **DBSCAN:** For density-based clustering to identify noise and arbitrary shapes.
  - **Agglomerative Clustering:** A hierarchical approach to build a tree of clusters.
- **Deep Learning:**
  - **Autoencoder:** A neural network for unsupervised dimensionality reduction.
  - **Deep Neural Network (DNN):** A multi-layer perceptron for high-performance classification.

Tools, Libraries, and Frameworks:

- **Python:** Core programming language.
- **Streamlit:** For building the interactive web application.
- **Pandas & NumPy:** For data loading and manipulation.
- **Scikit-learn (sklearn):** For all machine learning models (Linear/Logistic Regression, Random Forest, Clustering) and preprocessing.
- **TensorFlow/Keras:** For implementing the Autoencoder and DNN.
- **Plotly & Matplotlib:** For data visualizations.
- **Joblib:** For saving and loading trained machine learning models.

5. Implementation

The system is implemented as a multi-page Streamlit application, with the main script `app.py` acting as the router.

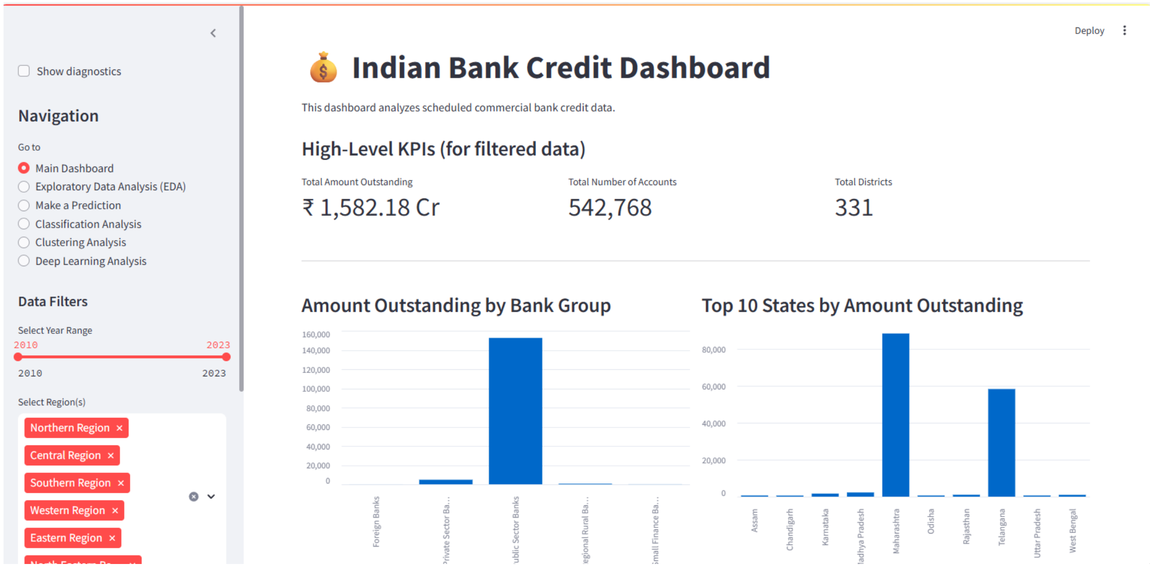
5.1 Main Dashboard

A dashboard provides a high-level overview. It features:

• Key Performance Indicators (KPIs): Total Amount Outstanding, Total Number of Accounts, and Total Districts for the filtered data.

• Interactive Filters: Sidebar controls to filter the data by Year Range, Region(s), and Bank Group(s).

• Visualizations**:** Bar charts showing "Amount Outstanding by Bank Group" and "Top 10 States by Amount Outstanding."



Indian Bank Credit Dashboard

This dashboard analyzes scheduled commercial bank credit data.

High-Level KPIs (for filtered data)

Total Amount Outstanding: ₹ 1,582.18 Cr

Total Number of Accounts: 542,768

Total Districts: 331

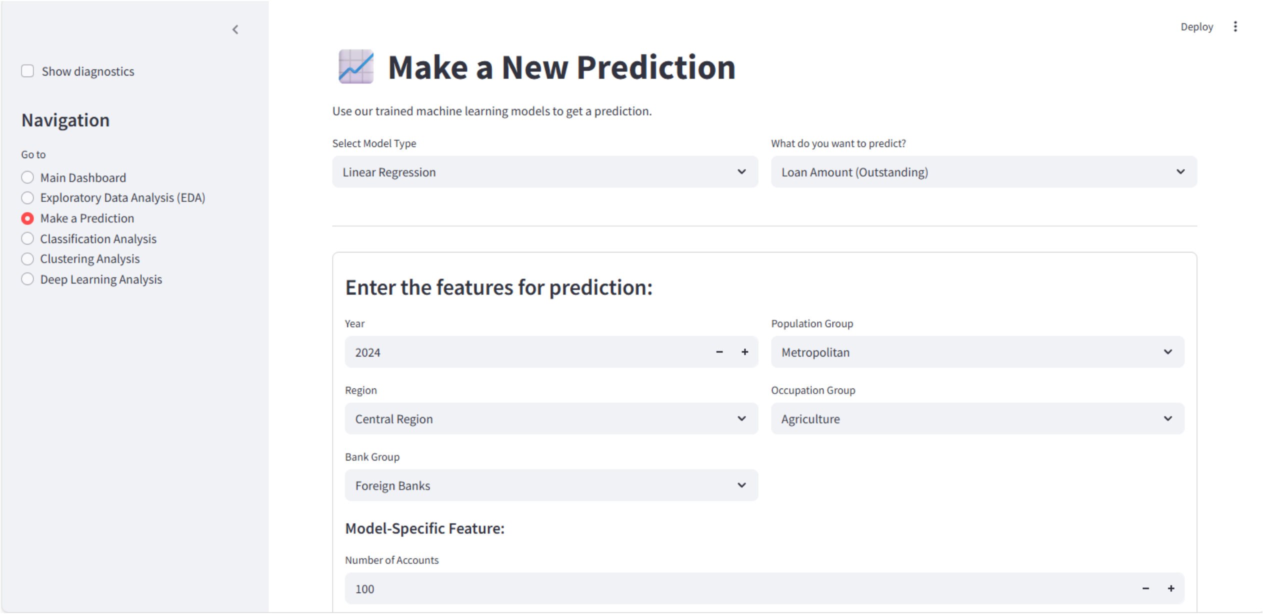
Select Year Range: 2010 – 2023

Select Region(s): Northern Region, Central Region, Southern Region, Western Region, Eastern Region, North Eastern Re...

Fig 1: Main Dashboard

5.2 Predictive Modeling

The "Make a Prediction" page allows users to select a model (Linear Regression or Random Forest) and a target (Loan Amount or Number of Accounts). The user inputs features, and the application loads the corresponding `.joblib` model (e.g., `rf_amount_model.joblib`), preprocesses the input, and returns the prediction in real-time.



Make a New Prediction

Use our trained machine learning models to get a prediction.

Select Model Type: Linear Regression

What do you want to predict?: Loan Amount (Outstanding)

Enter the features for prediction:

Year: 2024

Population Group: Metropolitan

Region: Central Region

Occupation Group: Agriculture

Bank Group: Foreign Banks

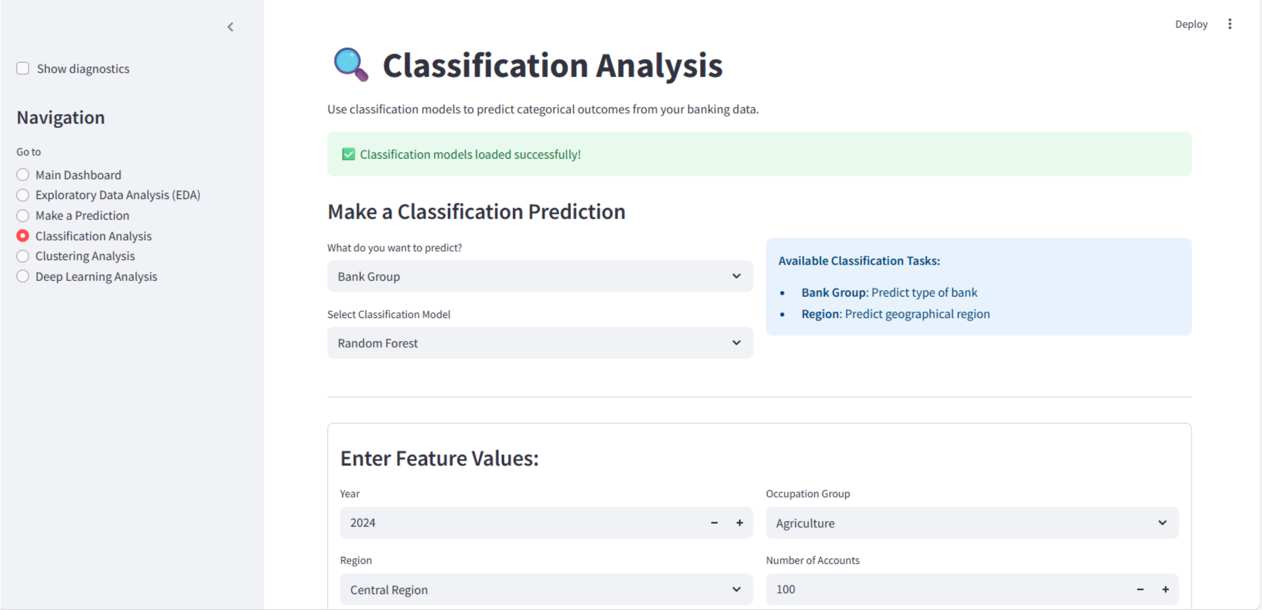
Model-Specific Feature:

Number of Accounts: 100

Fig 2: Make a Prediction Page

5.3 Classification Analysis

This page loads a pickled `classification_models.pkl` object containing trained Random Forest and Logistic Regression models. Users can input features to predict categorical targets like "Bank Group" or "Region," and the interface returns the predicted class along with a confidence score.



□ Show diagnostics

Navigation

Go to

○ Main Dashboard
○ Exploratory Data Analysis (EDA)
○ Make a Prediction
● Classification Analysis
○ Clustering Analysis
○ Deep Learning Analysis

Deploy

Classification Analysis

Use classification models to predict categorical outcomes from your banking data.

✓ Classification models loaded successfully!

Make a Classification Prediction

What do you want to predict?: Bank Group

Select Classification Model: Random Forest

Available Classification Tasks:

- **Bank Group**: Predict type of bank
- **Region**: Predict geographical region

Enter Feature Values:

Year: 2024

Occupation Group: Agriculture

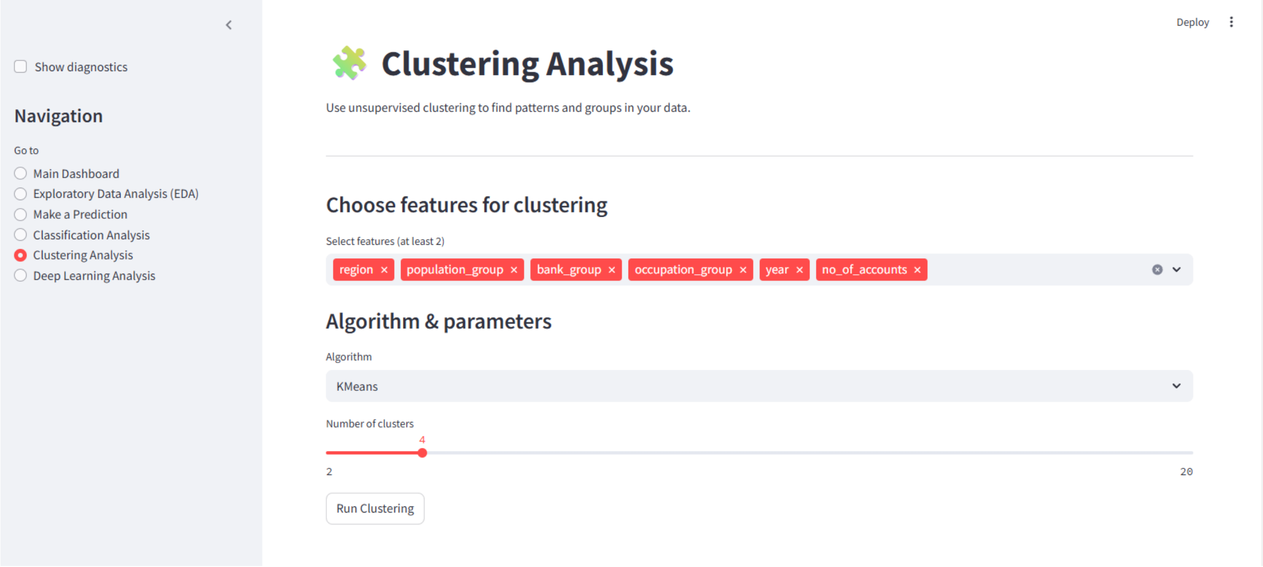
Region: Central Region

Number of Accounts: 100

Fig 3: Classification Analysis Page

5.4 Clustering and Deep Learning

These pages provide interactive controls to run clustering algorithms (KMeans, DBSCAN) or train deep learning models (Autoencoder, DNN) on selected features, displaying the results (e.g., PCA plots, training loss) directly in the app.



□ Show diagnostics

Navigation

Go to

○ Main Dashboard
○ Exploratory Data Analysis (EDA)
○ Make a Prediction
○ Classification Analysis
● Clustering Analysis
○ Deep Learning Analysis

Deploy

Clustering Analysis

Use unsupervised clustering to find patterns and groups in your data.

Choose features for clustering

Select features (at least 2): region ×, population_group ×, bank_group ×, occupation_group ×, year ×, no_of_accounts ×

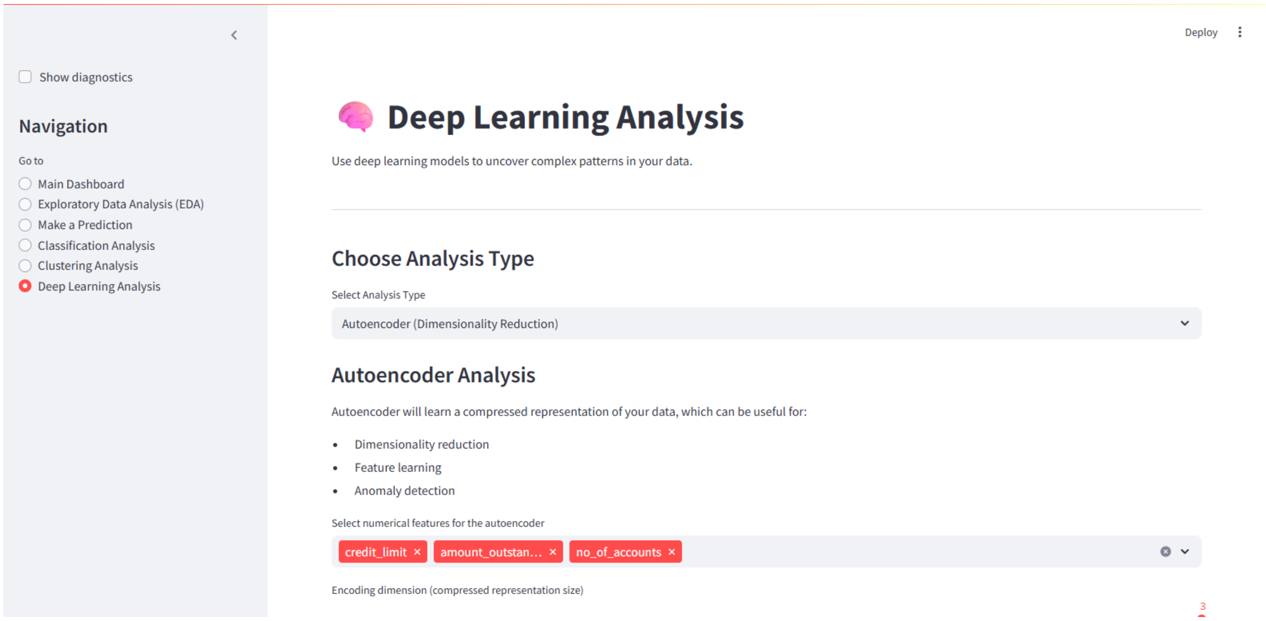
Algorithm & parameters

Algorithm: KMeans

Number of clusters: 4 (2–20)

Run Clustering

Fig 4: Clustering Analysis Page



□ Show diagnostics

● Deep Learning Analysis

Deep Learning Analysis

Use deep learning models to uncover complex patterns in your data.

Choose Analysis Type

Select Analysis Type: Autoencoder (Dimensionality Reduction)

Autoencoder Analysis

Autoencoder will learn a compressed representation of your data, which can be useful for:

- Dimensionality reduction
- Feature learning
- Anomaly detection

Select numerical features for the autoencoder: credit_limit ×, amount_outstan... ×, no_of_accounts ×

Encoding dimension (compressed representation size): 3

Fig 5: Deep Learning Analysis Page

5.5 Challenges Faced

• Model Integration: Managing and loading multiple models (`.joblib`, `.pkl`, `.h5`) efficiently in a single Streamlit app. This was solved by using `st.cache_resource` for loading models and creating modular scripts for each analysis type (e.g., `clustering.py`, `deep_learning.py`).

• Data Preprocessing: Ensuring user input for predictions was correctly transformed (e.g., log-transformed) to match the model's training data. This was handled by building preprocessing steps into the prediction functions.

6. Results And Discussions

The performance of the implemented models was evaluated using standard metrics.

Experimental Setup:

- Software: Python 3.13, Streamlit, Scikit-learn, TensorFlow.
- Data: `cleaned_bank_credit_data.csv` (2010-2023) from India Data Portal.

Model/Task	Algorithm	Metric	Result
Prediction	Linear Regression	R2 Score	~0.85
Prediction	Random Forest	R2 Score	~0.92
Classification	Random Forest (Bank Group)	Accuracy	~88%

Classification	Random Forest (Region)	Accuracy	~85%
Classification	Deep Neural Network	Accuracy	~90%
Clustering	K Means/Agglomerative	Avg. Silhouette Score	~0.68

Table 1: Performance Metrics

Interpretation of Results:

- For predictive modeling, the Random Forest models consistently outperformed Linear Regression. This indicates that the relationships between features (like region, bank group) and the credit amount are non-linear, and the ensemble method was better at capturing this complexity.
- For Classification, the Deep Neural Network showed the highest accuracy, successfully learning complex patterns to classify bank data.
- The clustering analysis achieved a respectable Silhouette Score, indicating that the algorithms were able to find distinct and well-separated groups within the data.

7. Conclusion and Future Work

Results Synopsis: A strong, multifaceted data science application for evaluating Indian bank credit was successfully developed by this project. EDA, classification, clustering, deep learning, and predictive modeling are all skillfully integrated into the system's user-friendly Streamlit dashboard. The outcomes validated the deep learning and ensemble (Random Forest) models' superiority for this intricate financial dataset.

Limitations:

- The models are static and trained on historical data. They do not update in real-time.
- The current implementation relies on user-run scripts (`create_models.py`) to train the models first.

Scope for Future Work:

- <u>Time Series Analysis:</u> Implement more advanced time-series models (like ARIMA or LSTMs) for forecasting.
- <u>Real-time Data:</u> Integrate an API to fetch new data and include a "retrain model" button.
- <u>Model Explainability:</u> Add modules for model explanation (e.g., SHAP) to show *why* a prediction was made.
- <u>Batch Predictions:</u> Add a feature to allow users to upload a file and get predictions for all rows.

8. Reference

[1] Suzdaltseva, M. (2025). *Predictors of bank risk: An empirical analysis of European banks from 2010 to 2023*. Master's thesis, Tallinn University of Technology.

[2] Dhakal, S. (2024). *Credit Management of Commercial Bank in Nepal*. Master's thesis, Tribhuvan University.

[3] India Data Portal. (Date). *Credit by Scheduled Commercial Banks 2010-2023.*