AI – ENHANCED SYNTHETIC CREDIT RISK ANALYSIS AND RAG-BASED FINANCIAL ASSISTANT

A project report submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE AND ENGINEERING

by

M.Dhanasri(23NM5A4207)
R.J.Harshitha(22NM1A4242)
D.Nihitha(22NM1A4213)
A.L.Sanjana(22NM1A4202)

Under the guidance of

Mrs.J.Nalini
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

VIGNAN'S INSTITUTE OF ENGINEERING FOR WOMEN (A)
(ApprovedbyAICTE,NewDelhi& AffiliatedtoJNTU- GV,Vizianagaram)Estd.2008
Kapujaggarajupeta, VSEZ (Post), Visakhapatnam – 530049.

2025

VIGNAN’S INSTITUTE OF ENGINEERING FOR WOMEN (A)

(ApprovedbyAICTE,NewDelhi& AffiliatedtoJNTU- GV,Vizianagaram)Estd.2008
Kapujaggarajupeta, VSEZ (Post), Visakhapatnam – 530049.

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the project report titled “ AI-ENHANCED SYNTHETIC CREDIT RISK ANALYSIS AND RAG BASED FINANCIAL ASSISTANT” is being submitted to the Department of CSE by M.Dhanasri (23NM5A4207), RJ.Harshitha(22NM1A4242),D.Nihitha(22NM1A4213),A.L.Sanjana (22NM1A4202), during the study of IV B.Tech I Semester of Bachelor of Technology in COMPUTER SCIENCE & ENGINEERING during the period December 2024 – April 2025 in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in COMPUTER SCIENCE & ENGINEERING. This project report has not been previously submitted to any other University/Institution for the award of any other degree

INTERNAL GUIDE

EXTERNAL EXAMINER-1

EXTERNALEXAMINER-2

DECLARATION

Weherebydeclarethatthisprojectentitled“AI–ENHANCED SYNTHETIC CREDIT RISKANALYSISANDRAG-BASEDFINANCIAL ASSISTANT” is the original work donebyusinpartialfulfillmentoftherequirementsfor the award of the degree of Bachelor ofTechnologyinComputerScience&Engineering, Jawaharlal Nehru Technological University,Vizianagaram.Thisprojectreporthasnotbeen previously submitted in any other university/Institutionfortheawardofanyotherdegree.

M.Dhanasri(23NM5A4207)
R.J.Harshitha(22NM1A4242)
D.Nihitha(22NM1A4213)
A.LSanjana(22NM1A4202)

ACKNOWLEDGEMENT

We are deeply grateful to our project guide, Mrs. J.Nalini, Assistant Professor, Department of Computer Science & Engineering, Vignan's Institute of Engineering for Women, whose valuable guidance and coordination were instrumental to the project's success.

Our heartfelt thanks to Dr. P. Vijaya Bharati, Professor & HOD, Department of Computer Science & Engineering, for providing the necessary resources and guidance throughout the project timeline.

We thank our esteemed Principal, Dr. B. Arundhati, for her support. We also appreciate the assistance provided by all faculty members during our project journey.

We express our sincere gratitude to everyone who supported us throughout this project. The experience has been both rewarding and educational.

M.Dhanasri
R.J.Harshitha
D.Nihitha
A.LSanjana

ABSTRACT

Inthe evolving landscape of financial risk management, the ability to simulate, analyze, andinterpret credit risk data is essential for building resilient systems. This project presents anAI-powered synthetic credit risk dashboard integrated with a Retrieval-Augmented Generation (RAG)-based financial assistant to enhance credit risk analysis, learning, and decision-making. The system first generates realistic synthetic data, including key risk parameters such as Probability of Default (PD), Loss Given Default (LGD), Exposure at Default (EAD), and Credit Ratings, allowing users to simulate diverse credit portfolios. This data feeds into an interactive dashboard built with Streamlit, where users can visualize risk distributions, detect trends, and download datasets for further modeling. Additionally, the solution integrates a RAG-based risk bot powered by Large Language Models (LLMs) and a vector database. The bot enables users to query financial terms, regulations (e.g., Basel III), and credit risk methodologies, offering accurate, documentgrounded responses. This dual setup — combining synthetic data generation with an intelligent Q&A assistant — bridges the gap between data-driven risk evaluation and domain knowledge acquisition.

Keywords : Python (Core Programming Language) Streamlit (Web Application Framework) Faker (Data Generation Library) LangChain (LLM Orchestration for RAG) OpenAI API / Bedrock API (Large Language Model services) FAISS (Vector Database for Semantic Search) Pandas (Data Manipulation and Analysis) NumPy (Numerical Computation) Matplotlib / Plotly (Optional – for custom visualizations)
AWS / Streamlit Cloud (Optional Deployment Platforms)

TABLE OF CONTENTS

INTRODUCTION

Inthemodern financial landscape, accurately assessing credit risk is vital for institutions to ensure stability, reduce loan defaults, and maintain compliance with regulatory frameworks. Traditional methods of risk analysis often depend heavily on historical data, manual analysis, and rigid models that struggle to adapt to dynamic market conditions.

This project introduces an AI-driven solution that not only generates synthetic credit risk data but also integrates a Retrieval-Augmented Generation (RAG)-based financial assistant. The solution leverages cutting-edge technologies like LangChain, Gemini, and FAISS to simulate real-world scenarios and assist users with intelligent financial queries.
The synthetic data generator built using Streamlit simulates values for key credit risk components such as:

o PD (Probability of Default)
o LGD (Loss Given Default)
o EAD (Exposure at Default)
o Credit Ratings

This synthetic data enables institutions to test models and visualize risks without compromising sensitive customer data. It is particularly useful for academic research, model prototyping, and testing in regulatory sandboxes.
To further enhance usability, the platform integrates a conversational assistant powered by Gemini, LangChain, and FAISS. This assistant retrieves relevant financial documents and provides contextual answers, improving decision-making and user engagement.

The goal of this project is to demonstrate how AI can make credit risk analysis more accessible, interactive, and scalable, while maintaining a focus on transparency and explainability. The system is designed to serve as both a simulation tool and an educational assistant for financial analysts, students, and fintech developers.

By combining synthetic data generation, intuitive visualization, and a smart RAG-based assistant, this project showcases the potential of generative AI in transforming traditional financial analysis into a dynamic, responsive, and user-friendly experience.

Traditional credit risk analysis often depends on large volumes of real customer data, which is sensitive and governed by strict privacy regulations. As a result, developing and testing new risk models becomes difficult for researchers and financial institutions that lack access to this data. Our approach solves this by generating high-quality synthetic data that mimics real credit risk profiles while ensuring data privacy.

The platform not only simulates important credit variables but also visualizes them through interactive dashboards built with Streamlit. These visualizations help users understand patterns, trends, and relationships in the data — like how changes in a borrower's financial behavior affect their credit rating or default probability.

On the AI side, the RAG-based chatbot assistant is designed to answer complex financial queries using document-based reasoning. It uses LangChain for orchestration, Gemini as the language model, and FAISS for efficient vector-based document retrieval. When a user asks a question, the system retrieves relevant information chunks from financial documents and responds with context-aware answers.

This assistant can serve a variety of use cases — such as explaining risk metrics, helping analysts navigate documentation, or answering regulatory questions. Unlike static dashboards, the assistant offers a conversational interface, making the analysis process more human-like and accessible even to non-technical users.

The integration of synthetic data generation with an intelligent assistant represents a novel approach to solving two critical problems: lack of training data and limited access to expert financial knowledge. With this tool, financial modeling becomes faster, safer, and smarter.

Overall, the project provides a cost-effective, secure, and scalable method for experimenting with risk models, training financial staff, and improving credit decision workflows. It sets a foundation for future applications in fintech, banking education, and risk management automation.

METHODOLOGY USED

1. Overall Approach :

- The project follows a modular and data-centric approach combining:
- Synthetic data generation for simulating financial credit risk variables.
- Interactive visualization using Streamlit.
- A smart, AI-powered RAG-based chatbot system to retrieve and respond to financial
- queries based on document context.
- Privacy by avoiding real user data.
- Scalability by modular design.
- Accessibility through interactive UI and conversational assistance.

2. Tools and Technologies Used :

Tool/Library	Purpose
Python	Core development language
Streamlit	Frontend UI and data dashboard framework
Faker	Generating realistic synthetic data
Matplotlib / Seaborn	Data visualization
LangChain	Retrieval-Augmented Generation (RAG) agent
FAISS	Vector database for document embedding search
HuggingFace Embeddings	To convert text into dense vectors
Gemini	LLM for intelligent response generation

3. Techniques Used :

- SyntheticDataSimulation:
  Credit parameters like Probability of Default (PD), Loss Given Default (LGD), Exposure at Default (EAD), and Ratings are generated using the Faker library and Python logic to simulate diverse financial profiles.
- DataVisualization:
  Charts (bar plots, histograms, etc.) are used to explore relationships and distribution of the synthetic data using Matplotlib and Seaborn.
- Retrieval-Augmented Generation (RAG):

□ Financial documents are split into text chunks using LangChain’s CharacterTextSplitter.

□ Thesechunksareconverted into embeddings using HuggingFace Transformers.

□ FAISSisusedtostore and retrieve relevant chunks based on query similarity.

□ Retrievedchunksare passed along with the query to Gemini, which generatesanaccurate and contextual response.

4. Workflow:

Thesystemfollowsamodular flow that ensures data privacy, usability, and intelligentassistance.Themajorworkflow steps are:

1.SyntheticDataGeneration:

Usertriggersdatacreation via the Streamlit UI.

PD,LGD,EAD,and Ratings are generated randomly and stored in a DataFrame.

2. Visualization:

Userselectsdesiredmetrics to view charts.

Chartsdisplaydistributions and trends of the generated data.

3.Document Preparation:

Financialtexts(e.g.,policy documents or explanations) are chunked.

Textchunksareembedded into vector form using HuggingFace embeddings.

4.QueryHandlingviaRAG:

Userasksafinancialquery through the chatbot interface.

Systemretrievesrelevant content from the FAISS vector database.

Retrievedcontentandthe user query are sent to Gemini for generating a contextualanswer.

5. Response Display:

Thefinalresponseisdisplayed in the chatbot interface.

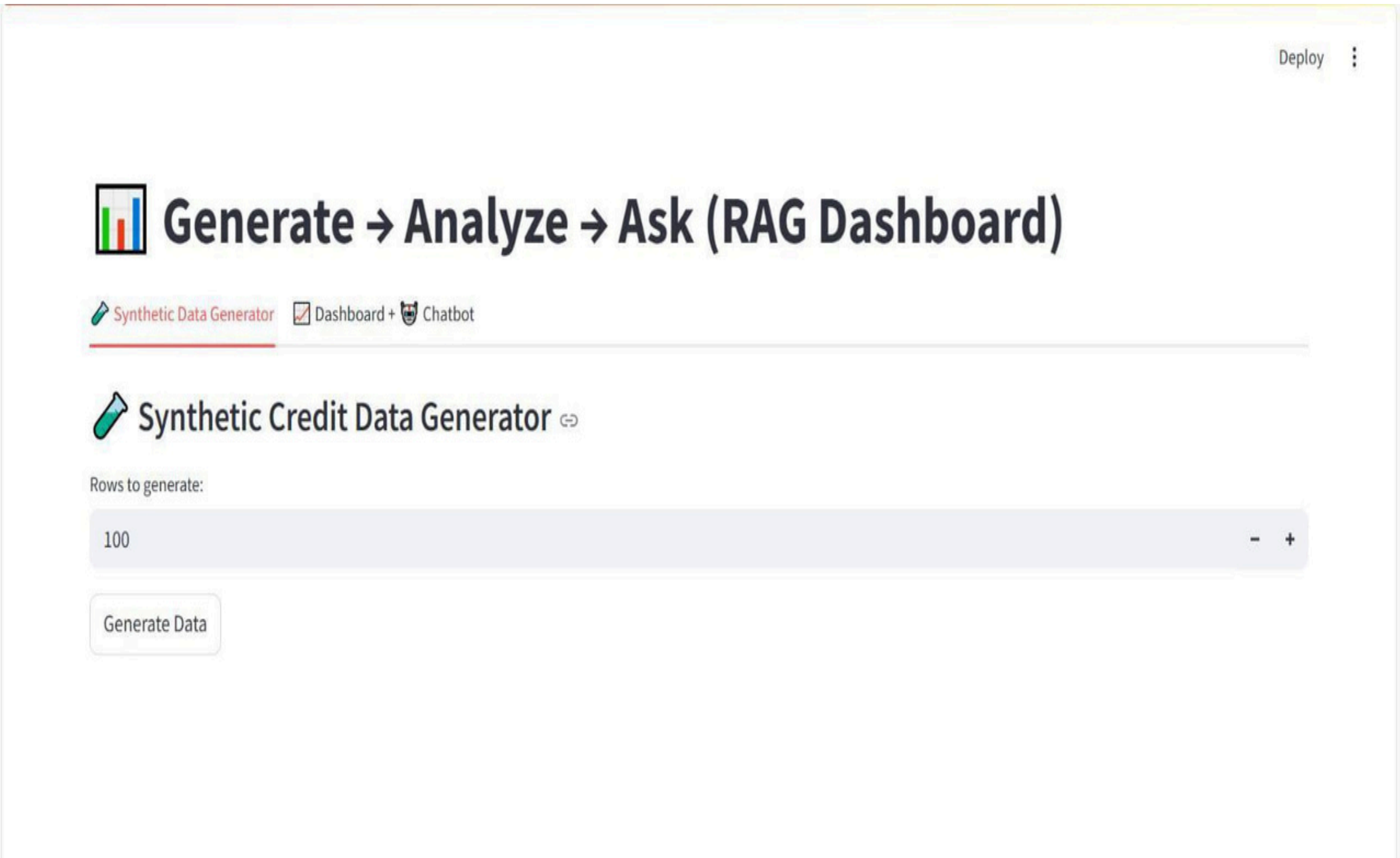
Thesystemmayinclude references or source text excerpts from the documentsused.

OUTCOMES AND PERCENTAGE OF WORK COMPLETED

1. Description of Results Achieved

The project has successfully achieved its intended goals by delivering an AI-powered credit risk platform that combines synthetic data simulation, risk visualization, and a smart conversational assistant. The main outcomes are:
A clean and interactive Streamlit UI, allowing users to generate synthetic credit data at the click of a button. The data covers key risk metrics like:

- Probability of Default (PD)
- Loss Given Default (LGD)
- Exposure at Default (EAD)
- Credit Rating



📊 Generate → Analyze → Ask (RAG Dashboard)

🧪 Synthetic Credit Data Generator

Rows to generate: 100

Image 6.1.1: Streamlit interface showing the generated synthetic credit risk data, including PD, LGD, EAD, and Ratings.

2. Integration of Interactive Data Visualization

    The application includes real-time visualizations that help users understand how different risk parameters behave. Users can choose metrics like PD (Probability of Default) or LGD (Loss Given Default) and generate corresponding bar charts or histograms to analyze the spread, trends, and concentration of risks across simulated credit profiles.

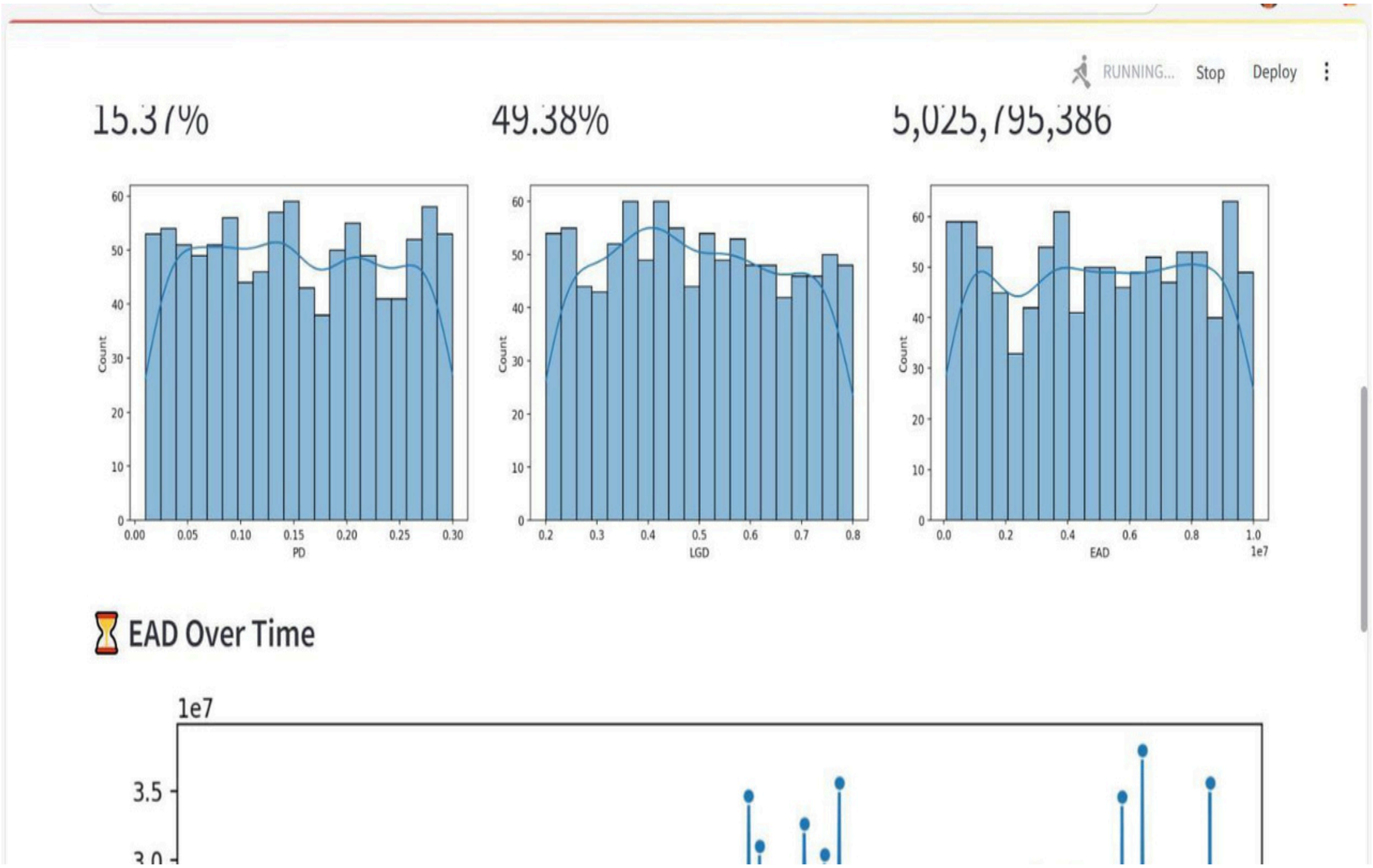
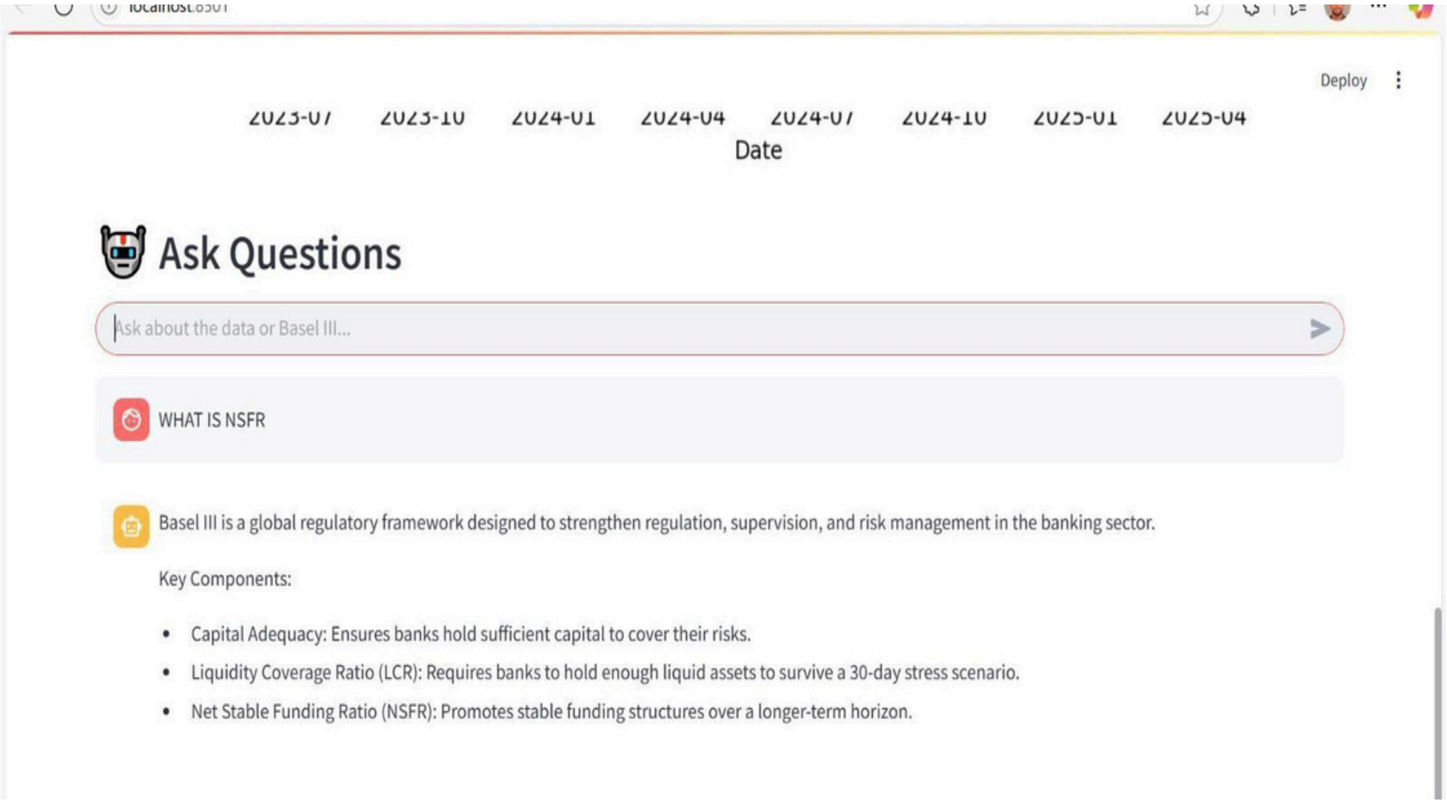


Image 6.1.2: bargraph showing the distribution of Probability of Default (PD) across customers.

3.Construction of a Retrieval-Augmented Generation (RAG) Chatbot System

A major outcome of the project is the development of a smart assistant using RAG architecture. When the user enters a query, the system retrieves the most relevant segments from embedded financial texts and forwards them to the Gemini LLM, which generates an accurate and contextual answer. This enables seamless document understanding for financial queries.



WHAT IS NSFR

Basel III is a global regulatory framework designed to strengthen regulation, supervision, and risk management in the banking sector.

Key Components:

- Capital Adequacy: Ensures banks hold sufficient capital to cover their risks.
- Liquidity Coverage Ratio (LCR): Requires banks to hold enough liquid assets to survive a 30-day stress scenario.
- Net Stable Funding Ratio (NSFR): Promotes stable funding structures over a longer-term horizon.

Image 6.1.3: Chatbot interface showing a contextual response to a financial query retrieved from document embeddings.

4.Estimated Completion Percentage

Module	Status	Completion %
Credit Risk Data Generation (Streamlit)	Completed	100%
Visualization of Risk Metrics	Completed	100%
RAG Assistant with LangChain & Gemini	Completed	100%
Vector Database Integration (FAISS)	Completed	100%
Testing, Validation & Documentation	Completed	100%

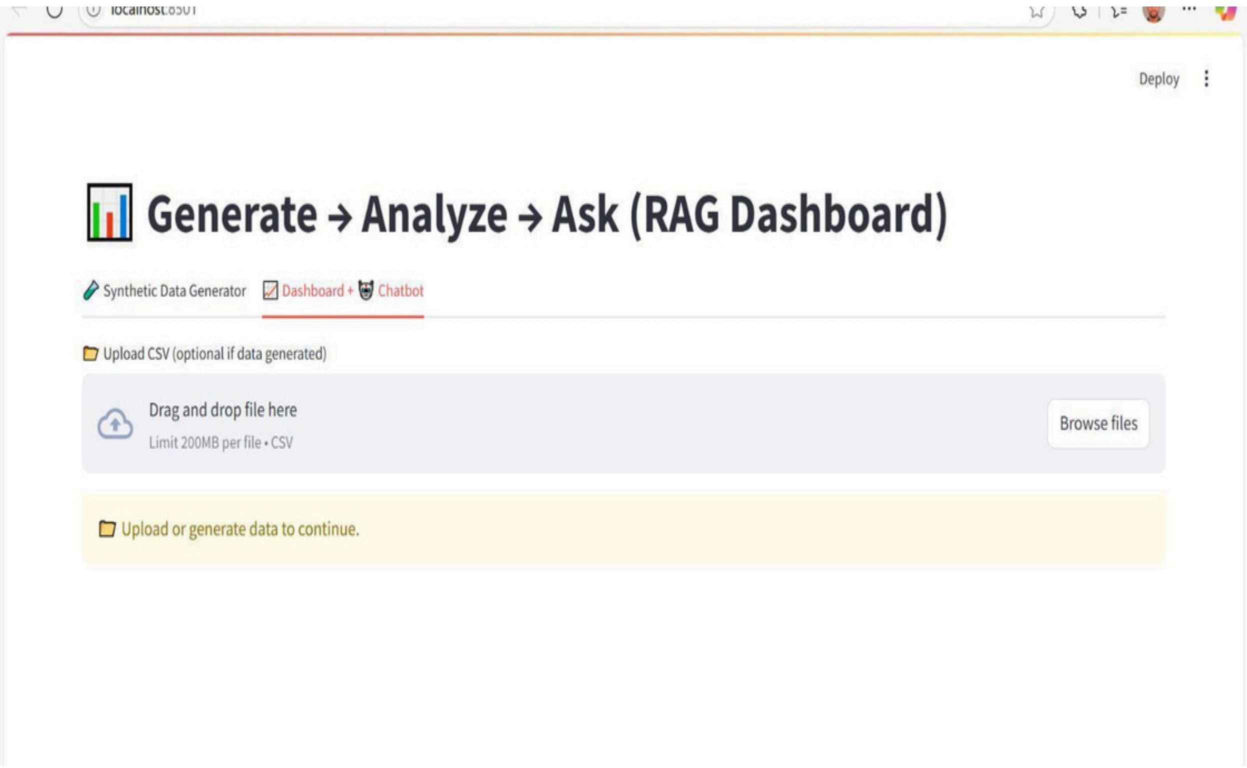
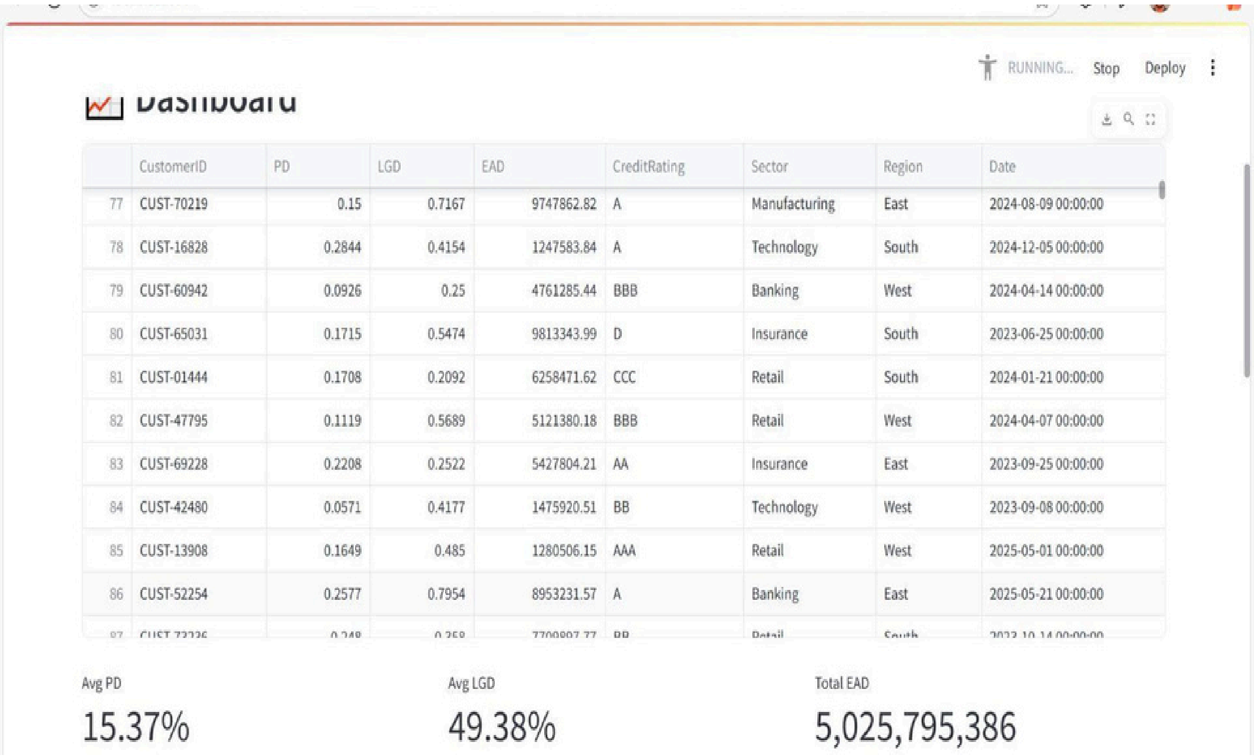

localhost:8501
Deploy
📊 Generate → Analyze → Ask (RAG Dashboard)
Synthetic Data Generator
Dashboard + Chatbot
Upload CSV (optional if data generated)
Drag and drop file here
Limit 200MB per file • CSV
Browse files
Upload or generate data to continue.

Image 6.1.3 basic interface



Dashboard

	CustomerID	PD	LGD	EAD	CreditRating	Sector	Region	Date
77	CUST-70219	0.15	0.7167	9747862.82	A	Manufacturing	East	2024-08-09 00:00:00
78	CUST-16828	0.2844	0.4154	1247583.84	A	Technology	South	2024-12-05 00:00:00
79	CUST-60942	0.0926	0.25	4761285.44	BBB	Banking	West	2024-04-14 00:00:00
80	CUST-65031	0.1715	0.5474	9813343.99	D	Insurance	South	2023-06-25 00:00:00
81	CUST-01444	0.1708	0.2092	6258471.62	CCC	Retail	South	2024-01-21 00:00:00
82	CUST-47795	0.1119	0.5689	5121380.18	BBB	Retail	West	2024-04-07 00:00:00
83	CUST-69228	0.2208	0.2522	5427804.21	AA	Insurance	East	2023-09-25 00:00:00
84	CUST-42480	0.0571	0.4177	1475920.51	BB	Technology	West	2023-09-08 00:00:00
85	CUST-13908	0.1649	0.485	1280506.15	AAA	Retail	West	2025-05-01 00:00:00
86	CUST-52254	0.2577	0.7954	8953231.57	A	Banking	East	2025-05-21 00:00:00
87	CUST-73236	0.248	0.358	7709807.77	BB	Retail	South	2023-10-14 00:00:00

Avg PD

15.37%

Avg LGD

49.38%

Total EAD

5,025,795,386

Image 6.1.4 Synthetic Data Generation

FUTURE SCOPE

Thecurrentsystem successfully demonstrates the generation of synthetic credit risk data and the implementation of a RAG-based intelligent financial assistant. However, there is significant potential to expand the platform further. Future enhancements may include:

Potential Enhancements

1. IntegrationwithMachineLearningModels

   Future versions can incorporate real-time credit scoring models (like Logistic Regression, Decision Trees, or XGBoost) trained on synthetic data to predict default probability or assign risk scores dynamically.

2. User-DrivenParameterControl

   Allowing users to set custom value ranges or probability distributions for synthetic data generation (e.g., defining expected PD range per customer type).

3. DynamicRiskAdjustment

   Introducing financial stress testing scenarios, such as economic downturn simulation, to observe how risk metrics like PD and EAD shift under stress.

4. InteractiveDashboards

   Enhancing visual analytics using interactive tools like Plotly or Dash for real-time filtering, grouping, and comparison of risk profiles.

Extensions

1. Multi-UserFunctionality

   Expanding the platform to support multiple users with login access, project saving, and individual result tracking.

2. DocumentUploadFeature

   Allowing users to upload their own PDFs or DOCs (like policy manuals or financial notes), which are embedded on the fly for chatbot interaction.

3. MultilingualSupport

   Adding translation and multilingual response capabilities to make the chatbot usable in regional and global financial institutions.

4. IntegrationwithFinancialAPIs

   Connecting the system with APIs like stock market feeds, loan processing systems, or regulatory data to make responses context-aware and up to date.

Future Directions

1. Cloud-BasedDeployment

   Hosting the application on platforms like AWS, Google Cloud, or Azure for secure, scalable, and real-time access.

2. RegulatoryComplianceSandbox

   Adapting the platform for testing models and strategies in environments simulating regulatory constraints (e.g., Basel III compliance).

3. EducationandTrainingUse

   Creating a training module using the system for finance students or credit analysts to learn about credit risk, synthetic data modeling, and AI-based assistants.

4. ExplainableAIIntegration

   Adding interpretability modules like SHAP or LIME to explain the decisions made by the chatbot or underlying models.

REFERENCES

- H. Sadok, F. Sakka, and M. E. H. El Maknouzi, Artificial Intelligence and Bank Credit Analysis: A Review, Financial Economics, 2021. Read the full article

- G. Namperumal, A. Selvaraj, and Y. Surampudi, Synthetic Data Generation for Credit Scoring Models: Leveraging AI and Machine Learning, Journal of Artificial Intelligence Research, vol. 2, no. 1, 2022. Explore the study

- K. Goyal, M. Garg, and S. Malik, Adoption of Artificial Intelligence-Based Credit Risk Assessment and Fraud Detection in Banking Services: A Hybrid Approach (SEM-ANN), Future Business Journal, vol. 11, 2025. View the paper

- R. Kumar Batchu, Artificial Intelligence in Credit Risk Assessment: Enhancing Accuracy and Efficiency, International Transactions in Artificial Intelligence, vol. 7, no. 7, 2023. Access the article

- Z. Chen and S. Kumar, Hybrid Synthetic Data in Banking Risk Models, Journal of Banking & Finance, vol. 124, 2021

🧠 RAG-Based Financial Assistant

- S. Kim et al., Optimizing Retrieval Strategies for Financial Question Answering Documents in Retrieval-Augmented Generation Systems, arXiv preprint arXiv:2503.15191, 2025. Check it out

- S. Zhao et al., FinRAGBench-V: A Benchmark for Multimodal RAG with Visual Citation in the Financial Domain, arXiv preprint arXiv:2505.17471, 2025. See the benchmark

- Sahil Sehgal, RAG-Based Financial Research Assistant for Amazon, GitHub Repository, 2025. Explore the implementation
- K. E. Kannammal et al., Fin-RAG: A RAG System for Financial Documents, International Journal of Innovative Science and Research Technology, vol. 10, no. 4, 2025. Read the publication

CONCLUSION

Yes, I intend to pursue the AI-Enhanced Synthetic Credit Risk Analysis and RAG-Based Financial Assistant

The proposed system offers a modern, intelligent, and practical approach to financial risk assessment and decision support. By combining synthetic credit data generation with Retrieval-Augmented Generation (RAG) techniques powered by FAISS and Gemini, the system not only enables dynamic data analysis but also facilitates interactive financial guidance through a conversational interface. The use of tools such as Pandas, Seaborn, Streamlit, and LangChain ensures transparency, scalability, and ease of implementation.

The modular design of the system allows for flexibility in both the data simulation process and the chatbot's domain knowledge, making it suitable for educational, training, and real-world financial applications. The project demonstrates key capabilities in data engineering, machine learning, vector-based search, and AI-powered language generation. With its interactive dashboard, data visualizations, and smart assistant, the solution bridges the gap between complex financial data and user-friendly insights.

This project marks a meaningful advancement beyond my previous mini-project, offering greater technical depth in natural language processing, prompt engineering, synthetic data pipelines, and real-time AI interaction. Its relevance to current trends in financial technology and AI applications makes it a strong representation of my growing expertise in intelligent systems and applied data science.