Synthetic Data Generation and Prediction for Corporate Rating using Bayesian and MCMC Techniques

Ayushman Anupam (MDS202411)
github.com/AyushmanGHub

Supervisor: Dr. M. R. Srinivasan
Chennai Mathematical Institute

Data Science Programme
Chennai Mathematical Institute

Contents

1 Abstract

Real-world corporate credit rating datasets are often limited, confidential, and highly imbalanced, making the development of stable predictive models difficult. Classical machine learning approaches struggle in this environment because they require complete retraining when new data arrives and provide only point estimates without uncertainty quantification—an essential requirement in financial risk analysis.

This project presents a Bayesian–MCMC based framework that integrates synthetic data generation, probabilistic credit rating prediction, and incremental Bayesian updating. A hybrid generative pipeline—combining Gaussian Mixture Models (GMM), quantile alignment, and an MCMC-based refinement step—produces high-fidelity synthetic datasets that closely match real data distributions.

For rating prediction, Bayesian logistic and multinomial regression models are constructed, and posterior samples are obtained using the Metropolis–Hastings algorithm. These samples enable full uncertainty quantification through credible intervals. A GMM-based residual correction step further stabilizes these predictions. The proposed framework outperforms classical baseline models for binary rating classification on both real and synthetic datasets while offering transparent and uncertainty-aware outputs.

Overall, this system provides a data-efficient, explainable, and industry-ready approach to corporate credit rating, with strong applicability in banking, credit rating agencies, NBFCs, and investment research environments.

2 Introduction

Corporate credit ratings play a critical role in financial decision-making, influencing lending policies, investment strategies, and regulatory compliance. However, real-world credit rating datasets are typically limited, confidential, and often imbalanced, making it challenging to build robust predictive models. Traditional machine learning approaches face additional constraints: they require complete retraining when new observations arrive, struggle with noisy or missing financial ratios, and provide only deterministic outputs without quantifying uncertainty. These limitations create significant barriers for institutions that rely on reliable, explainable, and adaptive risk assessment systems.

Bayesian methods, combined with Markov Chain Monte Carlo (MCMC) techniques, offer a principled solution to many of these challenges. By treating model parameters as probability distributions rather than fixed values, Bayesian models naturally incorporate uncertainty, enable incremental updating, and provide more interpretable predictions. At the same time, synthetic data generation has emerged as a practical approach to overcome data scarcity and confidentiality restrictions in credit risk modeling.

In this project, I develop a comprehensive Bayesian-MCMC framework for both synthetic data generation and corporate credit rating prediction. The generative pipeline employs a hybrid Gaussian Mixture Model (GMM), quantile alignment, and MCMC-based refinement to create synthetic datasets that closely replicate the statistical properties of real financial data. For predictive modeling, Bayesian logistic and multinomial regression models are constructed, and posterior samples are obtained using the Metropolis-Hastings algorithm. These posterior draws enable full uncertainty quantification through credible intervals, while a residual-based GMM correction further stabilizes predictive performance.

The overall objective of this work is to design a data-efficient, uncertainty-aware, and industry-aligned credit rating system that addresses key limitations in existing approaches. Experimental results demonstrate that the proposed framework enhances prediction accuracy, improves robustness to data noise and imbalance, and provides interpretable uncertainty estimates—making it suitable for banks, credit rating agencies, NBFCs, and investment research environments.

3 Problem Statement

Corporate credit rating prediction is a critical task for financial institutions, yet real-world rating data presents several challenges that limit the effectiveness of traditional modeling approaches. First, credit rating datasets are often highly confidential and sparsely available, restricting the development and validation of robust machine learning models. Existing public datasets are small, imbalanced across rating classes, and contain noisy or inconsistent financial ratios.

Second, classical machine learning models must be fully retrained whenever new financial data arrives. This makes them computationally expensive, slow to update, and unsuitable for environments where financial statements arrive periodically. Moreover, deterministic models provide only point predictions and fail to quantify uncertainty—an essential requirement for risk-sensitive applications such as lending, investment decisions, and regulatory reporting.

Third, real-world financial data contains latent structures, hidden patterns, and multi-modal distributions that are not adequately captured by conventional models. Overfitting becomes a major concern when dealing with limited and high-dimensional datasets, reducing the reliability of predictions.

The key problems addressed in this work are therefore:

- **Data Scarcity and Confidentiality:** Limited real credit rating data restricts model performance and generalizability.
- **Class Imbalance and Noisy Ratios:** Highly skewed rating distributions and noisy financial features reduce model stability.
- **Lack of Uncertainty Estimation:** Classical ML models cannot express prediction uncertainty, which is crucial in financial risk modeling.
- **Expensive Retraining Requirements:** Existing models need complete retraining with every new batch of data.
- **Hidden Latent Structure:** Standard deterministic models fail to capture multi-modal and latent relationships in financial ratios.

Addressing these issues requires a modeling approach that is data-efficient, uncertainty-aware, and capable of handling the inherent complexity of financial data—motivating the development of a Bayesian-MCMC based synthetic data generation and credit rating prediction framework.

4 Proposed Solution

To address the challenges of data scarcity, class imbalance, uncertainty estimation, and the computational limitations of classical machine learning models, I used a comprehensive Bayesian-MCMC based framework for both synthetic data generation and corporate credit rating prediction.

The core idea behind my solution is to combine probabilistic modeling with generative methods so that the system remains data-efficient, uncertainty-aware. The proposed approach consists of three major components:

1. **Synthetic Data Generation:** A hybrid generative pipeline that uses a Gaussian Mixture Model (GMM) to capture the joint distribution of financial ratios and categorical variables. After initial sampling, quantile alignment and an MCMC-based refinement step is applied to correct marginal distributions, variance structure, and correlations. This generates high-fidelity synthetic datasets that preserve the statistical characteristics of real credit rating data while avoiding confidentiality restrictions.

2. **Bayesian Predictive Modeling:** To improve prediction stability and incorporate uncertainty, Bayesian logistic regression for binary ratings and Bayesian multinomial regression for the five-class rating problem is applied. Metropolis-Hastings sampler is applied, to obtain posterior distributions over model parameters rather than single-point estimates. These posterior samples enable uncertainty quantification through credible intervals and provide more interpretable and transparent predictions.

3. **Residual-Based Correction and Incremental Updating:** To further enhance predictive accuracy, a GMM model is fitted on residual distributions and is used to refine test-time probabilities. Additionally, the Bayesian framework allows it to perform incremental updates when new data arrives, eliminating the need to retrain the model from scratch.

Overall, the proposed solution provides a robust, flexible, and interpretable approach to corporate credit rating prediction. By generating synthetic data, modeling uncertainty, and enabling efficient updates, this project directly addresses the limitations of traditional machine learning models and aligns it with the practical needs of banks, credit rating agencies, NBFCs, and other financial institutions.

5 About the Dataset

The dataset used in this project is sourced from the publicly available *Corporate Credit Rating with Financial Ratios* dataset on Kaggle. It contains financial and rating information for a large set of corporate entities across multiple sectors and rating agencies. The dataset consists of **7805 observations** and **25 features**, covering both quantitative financial ratios and qualitative categorical attributes.

Key Characteristics of the Dataset:

- **Size:** 7805 rows and 25 columns.
- **Missing Values:** No missing values are present.
- **Attributes:** The dataset includes financial ratios, company identifiers, sectoral information, and credit ratings assigned by different agencies.
- **Target Variables:**
  - **Rating (Full Rating Scale)** – Original letter-grade ratings such as AAA, AA+, A, BBB, BB, B, etc.
  - **Binary Rating** – A simplified binary label indicating whether a company is *Investment Grade* or *Non-Investment Grade.*

Key Columns Include:

- Company identifiers: Rating Agency, Rating, Rating Date, CIK, Ticker.
- Categorical descriptors: SIC Code, Sector, Rating_Year.
- Financial ratios: Current Ratio, Debt/Equity Ratio, Gross Margin, Operating Margin, EBIT Margin, EBITDA Margin, ROE, ROA, ROI, Asset Turnover, Operating Cash Flow per Share, Free Cash Flow per Share, etc.

Label Transformation: To simplify the multiclass prediction problem, I regroup the original letter-grade ratings into a **five-class rating system** (AAA-AA, A, BBB, BB, B & Below). This transformation helps reduce class imbalance and aligns the ratings into broader, industry-meaningful categories.

This dataset provides a realistic foundation for modeling creditworthiness but also exhibits challenges such as class imbalance, multi-agency inconsistency, and complex distributional patterns—motivating the need for Bayesian modeling and synthetic data generation.

6 Data Preparation

To prepare the dataset for Bayesian modeling and synthetic data generation, these preprocessing steps including column removal, categorical encoding, and target transformation are applied.

Column Filtering

Non-informative fields such as company identifiers and full date strings are removed: `Corporation`, `Ticker`, `CIK`, and `Rating Date`. Since the rating year is extracted separately, the date column is no longer required.

Categorical Encoding

Categorical attributes are converted using one-hot encoding for: `SIC Code`, `Sector`, `Rating_Year`, and `Rating Agency`. This ensures all variables are compatible with Bayesian and linear models.

Target Transformation

To reduce class imbalance, original ratings are grouped into a **five-class system**: **AAA-AA**, **A**, **BBB**, **BB**, and **B & Below**, stored as `Rating_Class_5`. The `Binary Rating` variable is used for investment-grade vs. non-investment-grade classification.

Numerical Features

Financial ratios are standardized using $z = \frac{x-\mu}{\sigma}$ to stabilize computation and improve MCMC sampling efficiency.

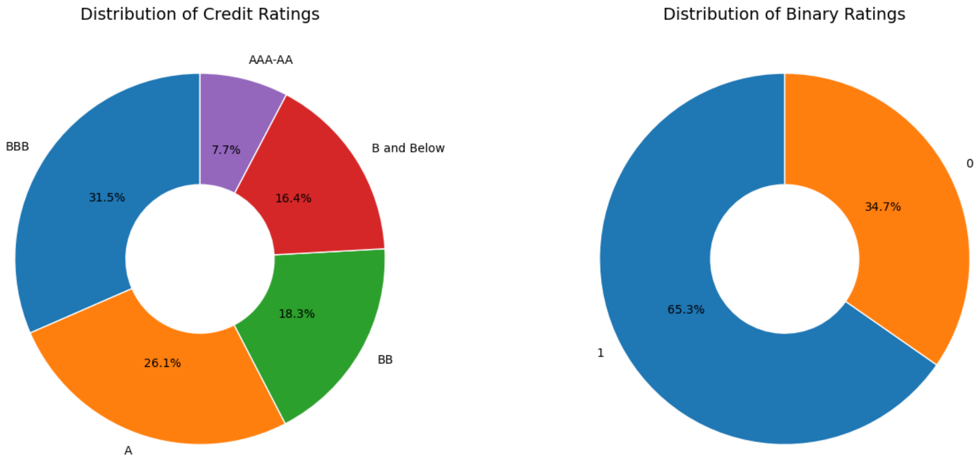
Final Dataset

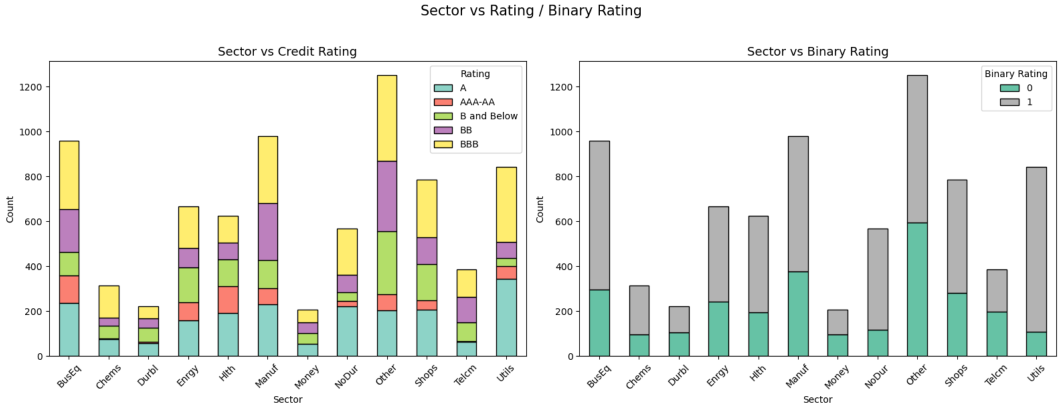
The processed dataset includes Standardized numeric features, one-hot encoded categorical variables, Two target variables, no missing values. This processed dataset provides a reliable basis for synthetic data generation and Bayesian MCMC modeling.

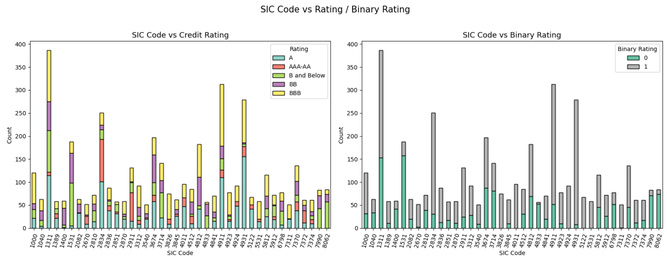
7 Exploratory Data Analysis

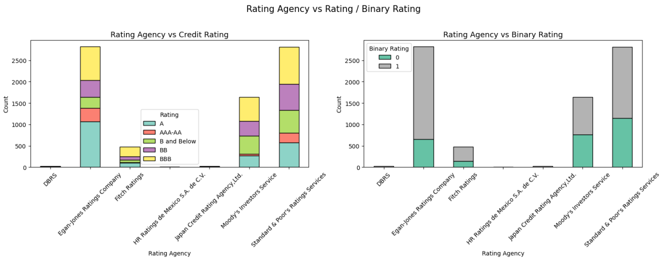
7.1 Target Distribution and Categorical Relationships

To understand the structure of the dataset, exploratory analysis is done, firstly target variable and its relationship with key categorical attributes such as Sector, SIC Code, and Rating Agency. These plots help identify class imbalance patterns and show how different industries and agencies contribute to the overall rating distribution.



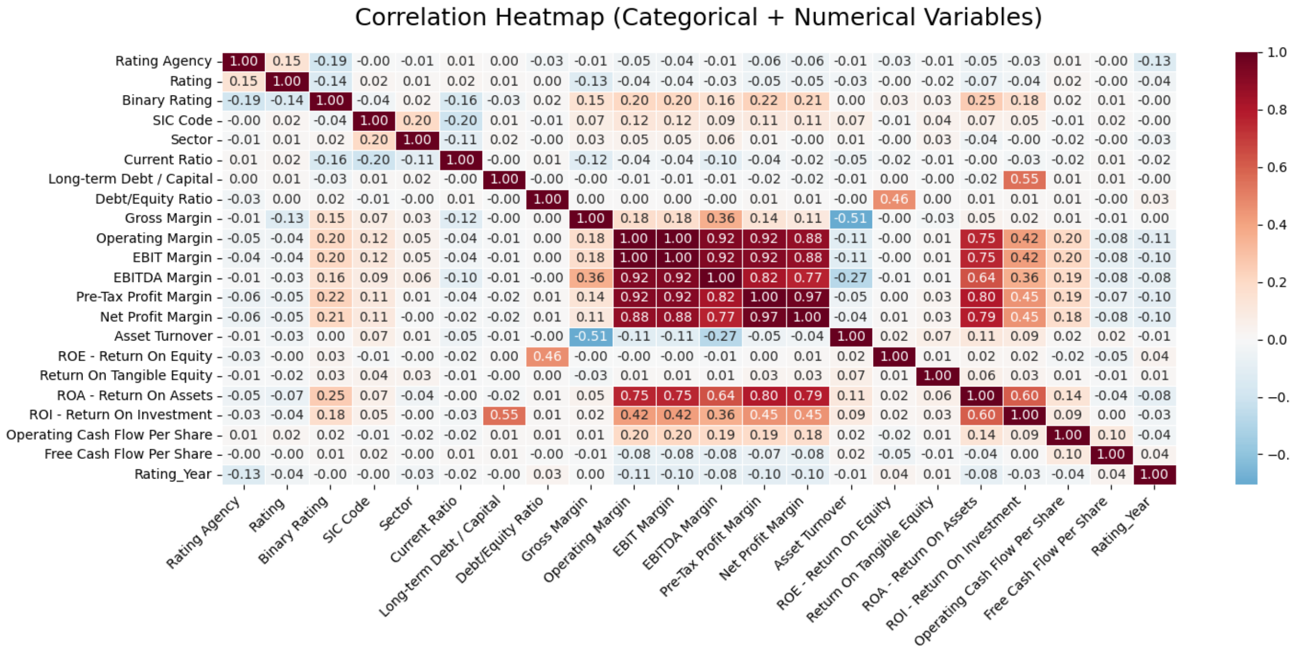






7.2 Correlation Heatmap

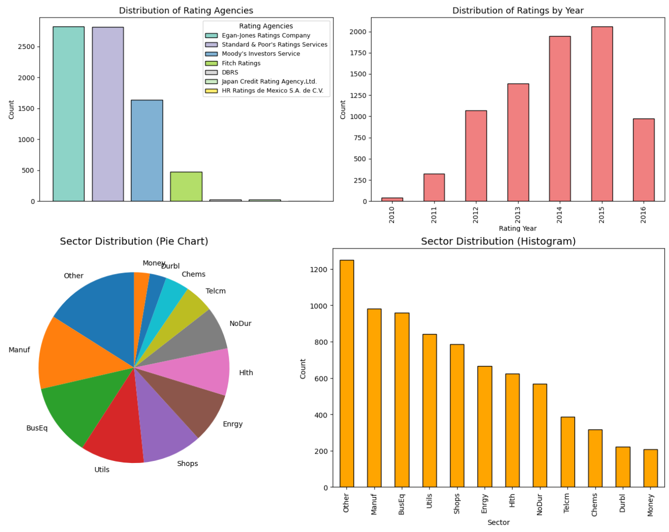
The correlation matrix highlights the relationships between numerical financial ratios. This provides insight into multicollinearity, which is important for both Bayesian modeling stability and for interpreting financial dependencies.



	Rating Agency	Rating	Binary Rating	SIC Code	Sector	Current Ratio	Long-term Debt / Capital	Debt/Equity Ratio	Gross Margin	Operating Margin	EBIT Margin	EBITDA Margin	Pre-Tax Profit Margin	Net Profit Margin	Asset Turnover	ROE - Return On Equity	Return On Tangible Equity	ROA - Return On Assets	ROI - Return On Investment	Operating Cash Flow Per Share	Free Cash Flow Per Share	Rating_Year
Rating Agency	1.00	0.15	-0.19	-0.00	-0.01	0.01	0.00	-0.03	-0.01	-0.05	-0.04	-0.01	-0.06	-0.06	-0.01	-0.03	-0.01	-0.05	-0.03	0.01	-0.00	-0.13
Rating	0.15	1.00	-0.14	0.02	0.01	0.02	0.01	0.00	-0.13	-0.04	-0.04	-0.03	-0.05	-0.05	-0.03	-0.00	-0.02	-0.07	-0.04	0.02	-0.00	-0.04
Binary Rating	-0.19	-0.14	1.00	-0.04	0.02	-0.16	-0.03	0.02	0.15	0.20	0.20	0.16	0.22	0.21	0.00	0.03	0.03	0.25	0.18	0.02	0.01	-0.00
SIC Code	-0.00	0.02	-0.04	1.00	0.20	-0.20	0.01	-0.01	0.07	0.12	0.12	0.09	0.11	0.11	0.07	-0.01	0.04	0.07	0.05	-0.01	0.02	-0.00
Sector	-0.01	0.01	0.02	0.20	1.00	-0.11	0.02	-0.00	0.03	0.05	0.05	0.06	0.01	-0.00	0.01	-0.00	0.03	-0.04	-0.00	-0.02	-0.00	-0.03
Current Ratio	0.01	0.02	-0.16	-0.20	-0.11	1.00	-0.00	0.01	-0.12	-0.04	-0.04	-0.10	-0.04	-0.02	-0.05	-0.02	-0.01	-0.00	-0.03	-0.02	0.01	-0.02
Long-term Debt / Capital	0.00	0.01	-0.03	0.01	0.02	-0.00	1.00	-0.00	-0.00	-0.01	-0.01	-0.01	-0.02	-0.02	-0.01	0.00	-0.00	-0.02	0.55	0.01	0.01	-0.00
Debt/Equity Ratio	-0.03	0.00	0.02	-0.01	-0.00	0.01	-0.00	1.00	0.00	0.00	0.00	-0.00	0.01	0.01	-0.00	0.46	0.00	0.01	0.01	0.01	-0.00	0.03
Gross Margin	-0.01	-0.13	0.15	0.07	0.03	-0.12	-0.00	0.00	1.00	0.18	0.18	0.36	0.14	0.11	-0.51	-0.00	-0.03	0.05	0.02	0.01	-0.01	0.00
Operating Margin	-0.05	-0.04	0.20	0.12	0.05	-0.04	-0.01	0.00	0.18	1.00	1.00	0.92	0.92	0.88	-0.11	-0.00	0.01	0.75	0.42	0.20	-0.08	-0.11
EBIT Margin	-0.04	-0.04	0.20	0.12	0.05	-0.04	-0.01	0.00	0.18	1.00	1.00	0.92	0.92	0.88	-0.11	-0.00	0.01	0.75	0.42	0.20	-0.08	-0.10
EBITDA Margin	-0.01	-0.03	0.16	0.09	0.06	-0.10	-0.01	-0.00	0.36	0.92	0.92	1.00	0.82	0.77	-0.27	-0.01	0.01	0.64	0.36	0.19	-0.08	-0.08
Pre-Tax Profit Margin	-0.06	-0.05	0.22	0.11	0.01	-0.04	-0.02	0.01	0.14	0.92	0.92	0.82	1.00	0.97	-0.05	0.00	0.03	0.80	0.45	0.19	-0.07	-0.10
Net Profit Margin	-0.06	-0.05	0.21	0.11	-0.00	-0.02	-0.02	0.01	0.11	0.88	0.88	0.77	0.97	1.00	-0.04	0.01	0.03	0.79	0.45	0.18	-0.08	-0.10
Asset Turnover	-0.01	-0.03	0.00	0.07	0.01	-0.05	-0.01	-0.00	-0.51	-0.11	-0.11	-0.27	-0.05	-0.04	1.00	0.02	0.07	0.11	0.09	0.02	0.02	-0.01
ROE - Return On Equity	-0.03	-0.00	0.03	-0.01	-0.00	-0.02	0.00	0.46	-0.00	-0.00	-0.00	-0.01	0.00	0.01	0.02	1.00	0.01	0.02	0.02	-0.02	-0.05	0.04
Return On Tangible Equity	-0.01	-0.02	0.03	0.04	0.03	-0.01	-0.00	0.00	-0.03	0.01	0.01	0.01	0.03	0.03	0.07	0.01	1.00	0.06	0.03	0.01	-0.01	0.01
ROA - Return On Assets	-0.05	-0.07	0.25	0.07	-0.04	-0.00	-0.02	0.01	0.05	0.75	0.75	0.64	0.80	0.79	0.11	0.02	0.06	1.00	0.60	0.14	-0.04	-0.08
ROI - Return On Investment	-0.03	-0.04	0.18	0.05	-0.00	-0.03	0.55	0.01	0.02	0.42	0.42	0.36	0.45	0.45	0.09	0.02	0.03	0.60	1.00	0.09	0.00	-0.03
Operating Cash Flow Per Share	0.01	0.02	0.02	-0.01	-0.02	-0.02	0.01	0.01	0.01	0.20	0.20	0.19	0.19	0.18	0.02	-0.02	0.01	0.14	0.09	1.00	0.10	-0.04
Free Cash Flow Per Share	-0.00	-0.00	0.01	0.02	-0.00	0.01	0.01	-0.00	-0.01	-0.08	-0.08	-0.08	-0.07	-0.08	0.02	-0.05	-0.01	-0.04	0.00	0.10	1.00	0.04
Rating_Year	-0.13	-0.04	-0.00	-0.00	-0.03	-0.02	-0.00	0.03	0.00	-0.11	-0.10	-0.08	-0.10	-0.10	-0.01	0.04	0.01	-0.08	-0.03	-0.04	0.04	1.00

7.3 Categorical Exploratory Data Analysis

Each categorical variable is analysed both visually and by checking the count of unique categories. This step clarifies dataset diversity and helps evaluate class sparsity in fields such as SIC Code and Rating Agency.

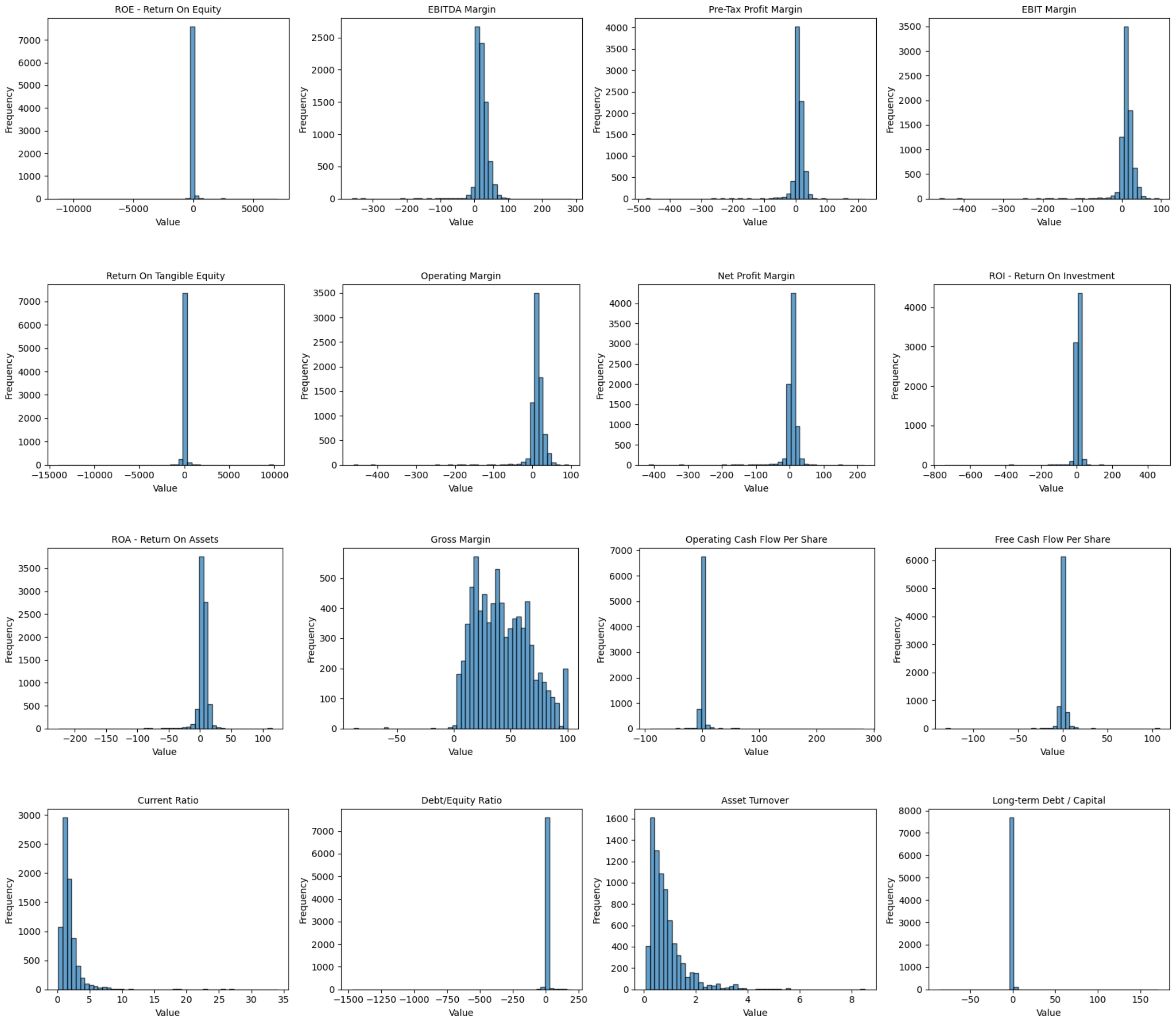


Unique values in each categorical field:

- SIC Code: **240**
- Sector: **12**
- Rating Agency: **7**
- Rating_Year: **7**
- Rating (5-class): **5**
- Binary Rating: **2**

7.4 Distribution of Numerical Features

To further explore numerical behaviour, histograms for all financial ratios are examined. These plots reveal skewness, heavy tails, multi-modal patterns, and outliers—properties that motivate the use of GMM + MCMC refinement in the synthetic data generation stage.



The exploratory analysis reveals clear class imbalance in both binary and multiclass rating targets, non-linear relationships across financial ratios, and high variability across sectors and rating agencies. These characteristics justify the use of Bayesian modeling, synthetic data generation, and MCMC techniques for robust prediction.

8 Synthetic Data Generation

To overcome data scarcity, confidentiality constraints, and class imbalance, Hybrid synthetic data generation pipeline combining Gaussian Mixture Modeling (GMM), quantile matching, and a lightweight MCMC refinement step is designed. This pipeline reconstructs the joint distribution of the original dataset and ensures that synthetic samples preserve both marginal behaviour and multivariate correlations.

8.1 GMM-Based Structural Modeling

Firstly, modeling of the dataset is done using a Gaussian Mixture Model (GMM) with component-wise handling of numerical and categorical variables. Numeric features are standardized, and the optimal number of mixture components is chosen using the Bayesian Information Criterion (BIC). For each GMM component, the following component estimate:

- Means and covariances for all numerical features,
- Probability tables for one-hot encoded and categorical attributes.

This allows generative model to capture multimodality, latent clusters, and the mixed data structure inherent in financial ratios and categorical descriptors.

8.2 Sampling Synthetic Observations

Once the GMM is trained, synthetic samples are generated in two steps:

- a mixture component is drawn according to GMM weights,
- numerical features from the corresponding multivariate Gaussian are sampled,
- categorical variables are sampled using component-level probability tables.

The results is an initial batch of synthetic observations that approximates the underlying cluster structure of the original data.

8.3 Distributional Refinement via Quantile Matching

Although GMM captures broad structure, marginal numerical distributions may still drift from the original data. Therefore, quantile alignment is applied:

- Each numerical column in the synthetic data is adjusted so that its empirical distribution matches the real one.
- One-hot inconsistencies are corrected post-alignment.

This step ensures distributional fidelity and reduces sampling distortions.

8.4 MCMC-Based Variance and Correlation Correction

To further enhance multivariate realism, a lightweight MCMC refinement step is used:

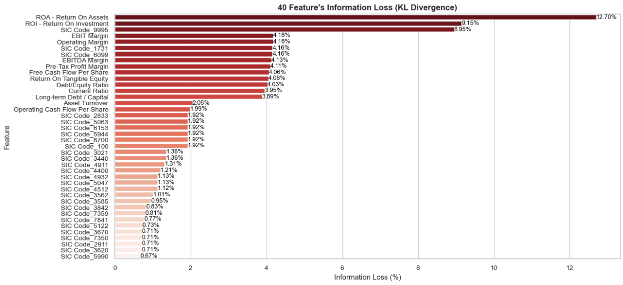
- Random perturbations are proposed for selected numeric dimensions,
- A loss function penalizes deviations in variance and correlation,
- Proposals are accepted using a Metropolis acceptance rule.

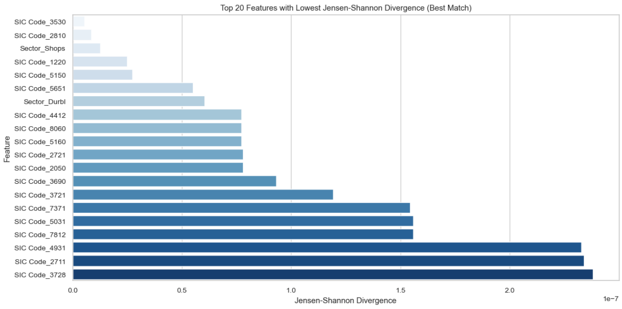
This process improves the covariance structure and better aligns key financial ratios such as ROA, ROI, and ROE with those in real data.

8.5 Quality Evaluation of Synthetic Data

Finally, to test similarity between real and synthetic data following metrics are used:

- **Kullback–Leibler Divergence (KL)** – measures information loss,
- **Jensen–Shannon Similarity (JSS)** – measures symmetric distributional closeness.





Across key financial ratios (ROA, ROI, leverage ratios, margins), KL divergence remains extremely low (0.002–0.007) and JSS exceeds **99%**, demonstrating that the synthetic data accurately preserves statistical properties of the original dataset.

The final synthetic dataset closely matches real financial behaviour at both marginal and multivariate levels. After the refinement step, the synthetic data is merged or used stand-alone for Bayesian modeling, enabling stable training, reduced overfitting, and data augmentation while preserving confidentiality.

9 Binary Rating Classification

To model investment–grade vs. non–investment–grade ratings, Bayesian logistic regression framework combined with Metropolis–Hastings MCMC sampling and a residual-based GMM refinement is used. This provides posterior uncertainty, probabilistic predictions, and improved prediction over classical logistic regression.

9.1 Bayesian Logistic Regression Model

The binary target variable $y_i \in \{0, 1\}$ is modeled with the standard logistic likelihood. Given predictor matrix X, the model is:

$$\log p(\beta \mid X, y) = \sum_i [y_i \log \sigma(X_i\beta) + (1 - y_i) \log(1 - \sigma(X_i\beta))] - \frac{1}{2}\beta^\top \Sigma^{-1} \beta,$$

where weak Gaussian $\beta \sim \mathcal{N}(0, 10I)$ is used as prior.

This formulation allows the posterior to balance between the likelihood and prior, reducing overfitting and capturing uncertainty in parameter estimates.

9.2 Metropolis–Hastings MCMC Sampling

To sample from the posterior distribution, random–walk Metropolis sampler is used:

- Initialize at $\beta^{(0)} = 0$,
- Propose $\beta' = \beta + \mathcal{N}(0, \text{proposal\_std}^2 I)$,
- Compute the log–acceptance ratio $\Delta = \log p(\beta') - \log p(\beta)$,
- Accept β' with probability $\min(1, e^{\Delta})$.

30,000 iterations were performed with a burn–in of **2000**, achieving stable acceptance and well–mixed posterior chains.

9.3 Posterior Predictive Probabilities and Credible Intervals

For each posterior sample $\beta^{(s)}$, the predicted probability is:

$$p_i^{(s)} = \sigma(X_i\beta^{(s)}).$$

And, point estimate for each sample is the posterior mean: $\hat{p}_i = \frac{1}{S}\sum_{s=1}^{S} p_i^{(s)}$.

A 95% credible interval is obtained using empirical percentiles of the predictive distribution $\{p_i^{(s)}\}$. This provides uncertainty-aware.

9.4 Residual-Based GMM Refinement

To improve predictive calibration, sample-wise residuals were computed

$$r_i^{(s)} = y_i - p_i^{(s)},$$

and were aggregated to obtain r_i. A small GMM (three components) was trained on the residual distribution.

For every test sample, posterior responsibilities γ_{ik} were calculated and a penalty was applied based on residual structure. This adjusted raw probabilities into refined probabilities $\tilde{p}_i$, improving classification stability.

9.5 Evaluation on Real and Synthetic Data

The final model was evaluated on both real and synthetic datasets.

Performance on Synthetic Data

- Accuracy: **0.8077**
- Precision: 0.8971
- Recall: 0.7981
- F1 Score: 0.8447

Sample Predictions (Top 5):

y	y_pred	raw	pen.	refined	CI (L,U)
1	1	0.8379	0.0980	0.7398	(0.6135, 0.8228)
1	0	0.3154	0.1548	0.1606	(-0.1508, 0.8056)
0	0	0.5493	0.2294	0.3198	(-0.1991, 0.7474)
0	1	0.8013	0.2497	0.5515	(0.0178, 0.7289)
1	1	0.7050	0.1133	0.5917	(0.3700, 0.7556)

Performance on Original Data

- Accuracy: **0.8949**
- Precision: 0.9388
- Recall: 0.9008
- F1 Score: 0.9194

Sample Predictions (Top 5):

y	y_pred	raw	pen.	refined	CI (L,U)
1	1	0.9068	0.0500	0.8568	(0.8201, 0.8861)
0	0	0.0802	0.0506	0.0296	(-0.0183, 0.0932)
1	1	0.9957	0.0053	0.9904	(0.9791, 0.9940)
0	0	0.2465	0.0735	0.1730	(0.0271, 0.3263)
1	1	0.9054	0.0199	0.8855	(0.4840, 0.9800)

As we can see, Bayesian model significantly improves binary rating prediction over the classical logistic regression baseline. Posterior samples provide credible intervals for uncertainty, while the GMM-based residual refinement enhances calibration. The model maintains strong performance even when trained using synthetic data, demonstrating the quality of the generated dataset and the robustness of the Bayesian framework.

10 Multiclass Rating Classification

To predict the full five-class corporate credit rating (AAA-AA, A, BBB, BB, B & Below), a Bayesian multinomial logistic regression model was used. This approach provides probabilistic predictions, posterior uncertainty, and avoids overfitting through prior regularization. Posterior inference is performed using a Metropolis-Hastings MCMC sampler.

10.1 Bayesian Multinomial Logistic Regression Model

For a classification task with K rating classes, parameter vectors is estimated for $(K-1)$ classes, using the final class as the reference category. Given predictors X_i, the linear score for class c is:

$$\eta_{i,c} = X_i B_c,$$

and the class probabilities are obtained using the softmax function:

$$p_{i,c} = \frac{\exp(\eta_{i,c})}{\sum_{j=1}^{K} \exp(\eta_{i,j})}.$$

To regularize the model and prevent overfitting, weak prior is placed:

$$B_c \sim \mathcal{N}(0, 10I).$$

This prior ensures stability when classes are imbalanced or highly correlated.

10.2 Posterior Sampling Using Metropolis-Hastings MCMC

The coefficient matrix B is estimated by sampling from its posterior distribution:

- Initialize $B^{(0)} = 0$,
- Propose a candidate matrix

$$B' = B + \mathcal{N}(0, \sigma^2 I),$$

- Compute the acceptance probability $\min\{1, \exp(\log p(B') - \log p(B))\}$,
- After burn-in, collect posterior draws $\{B^{(s)}\}_{s=1}^{S}$.

Each posterior sample corresponds to a different set of rating-class boundaries, capturing uncertainty inherent in the classification task.

10.3 Posterior Predictive Probabilities and Credible Intervals

For each test instance i, and for each posterior sample $B^{(s)}$, I compute the class probability vector: $p_{i,c}^{(s)} = \text{softmax}(X_i B^{(s)})_c$.
This produces a full posterior distribution over class probabilities. The final predicted probability for each class is the posterior mean: $\hat{p}_{i,c} = \frac{1}{S}\sum_{s=1}^{S} p_{i,c}^{(s)}$.

A **95% credible interval** is obtained from the 2.5th and 97.5th percentiles of the posterior distribution: $\text{CI}_{i,c} = \left(\text{Percentile}_{2.5}(p_{i,c}^{(s)}),\ \text{Percentile}_{97.5}(p_{i,c}^{(s)})\right)$. So, each prediction includes uncertainty across all rating classes.

10.4 Result

The following table shows selected predictions with posterior means and 95% credible intervals across all classes. Accuracy on synthetic real was **0.5452** and synthetic data was **0.4795**.

Results on Actual Data

Actual	Pred	A (L,P,U)	AAA-AA (L,P,U)	B&Below (L,P,U)	BB (L,P,U)	BBB (L,P,U)
BBB	A	(0.277,0.450,0.536)	(0.051,0.123,0.191)	(0.027,0.056,0.114)	(0.080,0.149,0.225)	(0.172,0.219,0.314)
B&Below	B&Below	(0.003,0.009,0.036)	(0.005,0.024,0.099)	(0.341,0.453,0.545)	(0.283,0.362,0.441)	(0.088,0.149,0.216)
A	A	(0.445,0.595,0.725)	(0.067,0.166,0.274)	(0.017,0.039,0.126)	(0.028,0.050,0.083)	(0.103,0.148,0.214)
B&Below	B&Below	(0.020,0.087,0.239)	(0.054,0.141,0.216)	(0.290,0.565,0.743)	(0.035,0.081,0.166)	(0.053,0.123,0.185)
BBB	BBB	(0.029,0.133,0.400)	(0.006,0.051,0.344)	(0.007,0.072,0.407)	(0.058,0.237,0.478)	(0.147,0.504,0.726)

Results on Synthetic Data

Actual	Pred	A (L,P,U)	AAA-AA (L,P,U)	B&Below (L,P,U)	BB (L,P,U)	BBB (L,P,U)
B&Below	B&Below	(0.126,0.173,0.219)	(0.031,0.053,0.101)	(0.220,0.298,0.352)	(0.162,0.237,0.319)	(0.160,0.236,0.332)
BB	BB	(0.117,0.181,0.240)	(0.058,0.144,0.267)	(0.042,0.097,0.170)	(0.207,0.296,0.438)	(0.190,0.280,0.387)
BBB	B&Below	(0.117,0.201,0.268)	(0.054,0.091,0.147)	(0.279,0.406,0.477)	(0.021,0.062,0.131)	(0.155,0.238,0.347)
A	B&Below	(0.063,0.238,0.453)	(0.025,0.082,0.235)	(0.096,0.285,0.536)	(0.060,0.144,0.229)	(0.159,0.247,0.310)
B&Below	B&Below	(0.000,0.012,0.119)	(0.000,0.013,0.053)	(0.234,0.543,0.814)	(0.124,0.350,0.649)	(0.009,0.080,0.197)

The Bayesian multinomial model produces full predictive probability distributions and credible intervals for each of the five rating classes. Although accuracy is lower than the binary model, this is primarily due to inconsistent rating standards across different credit rating agencies. The model remains valuable for uncertainty-aware multiclass credit risk assessment and demonstrates consistent behaviour on both real and synthetic datasets.

11 Base Model Evaluation

Before applying Bayesian and MCMC-based models, two classical machine learning are trained to act as a baseline for comparison: Logistic Regression for the binary rating task and Linear Discriminant Analysis (LDA) for the multiclass rating task. These provide a reference point to assess the improvement offered by the Bayesian framework.

11.1 Binary Rating Classification: Logistic Regression

The baseline model for the binary task (Investment Grade vs. Non-Investment Grade) is a standard logistic regression classifier. Its performance on both real and synthetic data is summarized below.

Performance on Original Data vs Synthetic Data

Original Data

- Accuracy: 0.8430
- Precision: 0.8589
- Recall: 0.9143
- F1 Score: 0.8857

Sample Predictions

ID	y_br	pred
4756	1	1
7379	0	0
6093	1	1
586	0	0
4791	1	1

Synthetic Data

- Accuracy: 0.7809
- Precision: 0.7955
- Recall: 0.8959
- F1 Score: 0.8427

Sample Predictions

ID	y_br	pred
107	0	0
5484	0	1
6998	1	1
3984	1	1
3111	0	0

11.2 Multi-Rating Classification: Linear Discriminant Analysis (LDA)

For the five-class rating task, I used a standard LDA classifier as the baseline model. This method provides a simple linear decision boundary across rating categories. Accuracy on original data was **0.5586** and on synthetuc data was **0.4850**. The performance on both real and synthetic datasets is presented below.

Performance on Original Data

ID	Actual	Pred
4756	BBB	A
7379	B&Below	B&Below
6093	A	A
586	B&Below	B&Below
4791	BBB	BBB

Performance on Synthetic Data

ID	Actual	Pred
107	B&Below	B&Below
5484	BB	BBB
6998	BBB	B&Below
3984	A	BBB
3111	B&Below	BB

Both baseline models perform reasonably well on original data but show reduced performance on synthetic data—particularly for the multiclass task. These results highlight the challenge of modelling complex rating boundaries using linear methods. This further motivates the Bayesian MCMC approach, which offers uncertainty quantification, better calibration, and improved performance in the binary classification setting.

12 Results and Model Comparison

This section compares the performance of the proposed Bayesian MCMC framework against classical baseline models for both binary and multiclass credit rating prediction. Results are reported on (i) the original dataset and (ii) the synthetic dataset generated using the GMM + quantile matching + MCMC refinement pipeline.

12.1 Comparison on Original Data

Binary Rating Classification

The Bayesian logistic regression with GMM-based residual refinement shows a consistent improvement across all four metrics, indicating better calibration and more stable classification boundaries.

Metric	Base Model (Logistic)	Proposed Model
Accuracy	0.8430	**0.8949**
Precision	0.8589	**0.9388**
Recall	0.9143	**0.9008**
F1 Score	0.8857	**0.9194**

Multiclass Rating Classification

The multiclass model achieves accuracy comparable to the LDA baseline. The slight drop is primarily due to inconsistent grading scales across rating agencies, which makes five-class classification inherently more challenging.

Metric	Base Model (LDA)	Proposed Model
Accuracy	0.5586	**0.5452**

12.2 Comparison on Synthetic Data

Binary Rating Classification

The proposed model again outperforms the baseline in accuracy, precision, and F1 score, demonstrating the usefulness of the synthetic dataset and the robustness of Bayesian inference.

Metric	Base Model (Logistic)	Proposed Model
Accuracy	0.7809	**0.8077**
Precision	0.7955	**0.8971**
Recall	0.8959	**0.7981**
F1 Score	0.8427	**0.8447**

Multiclass Rating Classification

Both models perform similarly on synthetic data. Since synthetic ratings are generated from multiple agencies with differing standards, maintaining fine-grained class boundaries is inherently difficult.

Metric	Base Model (LDA)	Proposed Model
Accuracy	0.4850	**0.4795**

12.3 Result Discussion

- **High-quality synthetic data** was generated using the hybrid GMM + quantile matching + MCMC refinement pipeline. KL divergence remained within **0.002–0.007** and Jensen–Shannon similarity exceeded **99%**, indicating excellent alignment between synthetic and real distributions.
- **Binary rating performance improved significantly**:
  - Accuracy improved from 0.8430 to **0.8949** on real data.
  - Accuracy improved from 0.7809 to **0.8077** on synthetic data.
  - Posterior distributions provided **credible intervals** for uncertainty-aware risk assessment.
- The **multiclass Bayesian model** produced complete posterior distributions and credible intervals for all five rating classes. However, its accuracy is affected by **agency-level inconsistencies** in rating scales. If ratings came from a single agency with a standardized scoring system, performance would likely improve substantially.

Overall, the Bayesian models—especially for the binary rating task—outperform classical baselines while offering uncertainty quantification, better calibration, and improved robustness on both real and synthetic datasets.

13 Conclusion

The results of this study demonstrate that Bayesian MCMC methods provide a reliable, transparent, and data-efficient framework for corporate credit rating prediction. Unlike classical machine learning models, which require large datasets and frequent retraining, the Bayesian approach naturally incorporates uncertainty, handles noisy or limited data, and supports incremental parameter updating as new observations become available.

A key contribution of this work is the development of a hybrid synthetic data generation pipeline based on Gaussian Mixture Models, quantile alignment, and MCMC distributional refinement. This pipeline effectively addresses the industry-wide challenge of data scarcity by producing statistically realistic synthetic datasets that closely match the distributional behaviour of real financial ratios. Low KL divergence and high Jensen–Shannon similarity scores confirm the high fidelity of the generated data.

The Bayesian logistic regression model with residual-based GMM correction shows clear improvements over the classical logistic regression baseline, particularly in binary rating prediction. The model provides posterior predictive intervals that quantify uncertainty, enhancing interpretability and supporting risk-sensitive decision making. The multiclass Bayesian model delivers full probability distributions and credible intervals across all five rating classes, offering deeper insight even though accuracy is affected by inconsistent rating standards across different agencies.

Overall, the proposed system presents a practical and robust solution for credit risk assessment. It is uncertainty-aware, adaptable to new data through Bayesian updating, and suitable for deployment in banking, credit rating agencies, NBFCs, investment research, and financial stress-testing frameworks. This work underscores the potential of Bayesian and MCMC-based techniques to enhance both predictive performance and model transparency in modern credit analytics.

14 Industry Challenges Addressed

Real-world credit risk modelling presents several practical challenges for financial institutions, banks, NBFCs, and credit rating agencies. These challenges limit the effectiveness of traditional machine learning models and motivate the need for more robust, probabilistic approaches. The key issues include:

- **Limited and confidential credit rating data**, which restricts the development, testing, and validation of accurate predictive models. Access to high-quality labelled datasets is often constrained due to confidentiality and regulatory policies.
- **Frequent updates to financial statements**, requiring models to be retrained regularly. Classical machine learning approaches must be fit from scratch each time, making them computationally expensive and slow to adapt.
- **Lack of uncertainty quantification** in deterministic models. Point predictions alone are insufficient for risk-sensitive decisions such as loan approval, portfolio allocation, risk provisioning, and regulatory reporting under frameworks like Basel and IFRS.
- **Financial ratios that are noisy, missing, or inconsistent** across companies and time periods, often leading to unstable models and overfitting.
- **Growing need for explainable and probabilistic forecasting**, as regulators and institutions increasingly demand transparent models capable of expressing confidence levels, stress scenarios, and uncertainty margins.

The Bayesian MCMC framework and synthetic data generation pipeline proposed in this work directly address these industry challenges by enabling uncertainty-aware predictions, incremental updating, and data augmentation using statistically realistic synthetic samples.

15 Industrial Application

The Bayesian–MCMC credit rating framework developed in this project directly addresses several real-world challenges faced by financial institutions, including data scarcity, expensive retraining cycles, and the need for transparent, uncertainty-aware predictions. Because the approach supports incremental learning, probabilistic model interpretation, and high-quality synthetic data generation, it is well-suited for deployment across multiple industry environments.

Potential Applications

- **Banks and NBFCs** The model can strengthen internal credit scoring systems, improve borrower risk assessment, and quantify the uncertainty in credit rating transitions. This is essential for underwriting, early-warning systems, and credit portfolio management.
- **Credit Rating Agencies** Synthetic datasets generated using GMM and MCMC refinement can support model development, stress testing, and regulatory audits without exposing sensitive proprietary financial information. This helps resolve the confidentiality barrier typically associated with corporate rating data.
- **Investment Firms and Asset Managers** Uncertainty-aware predictions can be incorporated into credit risk premia, portfolio rebalancing, and exposure management. Posterior intervals allow risk analysts to identify high-risk issuers and improve decision-making under uncertainty.
- **FinTech and Alternative Lending Platforms** Bayesian updating enables continuous model refinement as new borrower data arrives, avoiding costly full retraining. This makes the framework suitable for dynamic lending environments where quick adaptation is critical.

The flexibility, interpretability, and incremental learning capability of the proposed framework position it as a practical tool for modern credit risk analytics across regulated and data-constrained environments.

16 Future Work

This project demonstrates the potential of Bayesian modelling and synthetic data generation for improving corporate credit rating prediction. However, several promising directions remain open for future research and practical deployment.

- **Hierarchical Bayesian Models** Incorporating hierarchical priors can capture sector-level or industry-level structures, allowing different groups of firms to share information while maintaining individual characteristics. This would enhance model interpretability and reduce variance in data-scarce categories.
- **Advanced MCMC Techniques** The current framework uses Metropolis-Hastings sampling. More efficient methods such as Hamiltonian Monte Carlo (HMC) or No-U-Turn Sampler (NUTS) could improve convergence speed and posterior exploration in high-dimensional settings.
- **Synthetic Data Quality Improvements** Future work can explore:
  - Conditional synthetic generation (e.g., conditioning on sector or rating class),
  - Variational autoencoders (VAEs) or diffusion models for richer generative structure,
  - Fairness-aware synthetic sampling to handle biased rating distributions.
- **Unified Rating Standards** Since rating agencies use different grading scales, future models could integrate cross-agency mapping or standardization frameworks to reduce inconsistency in the multiclass prediction task.
- **Explainability and Regulatory Integration** Incorporating explainable AI (XAI) tools—such as Bayesian SHAP values, credible interval-based explanations, or posterior sensitivity analysis—would make the framework more suitable for Basel/IFRS-compliant decision processes.

Overall, future research can focus on scaling the Bayesian–MCMC framework to more complex datasets, improving generative quality, and integrating temporal and regulatory considerations to further strengthen its applicability in real-world credit risk modeling.

17 Code

This project is organized into a sequence of Jupyter notebooks, each handling a specific stage of the workflow. A brief description of the major notebooks is provided below:

- **01_DataPreparation.ipynb** Performs data cleaning, preprocessing, standardization, and encoding of both numerical and categorical variables.
- **02_DataGeneration-MCMC.ipynb** Implements Bayesian sampling and MCMC-based synthetic data generation.
- **03_EDA.ipynb** Conducts exploratory data analysis, summary statistics, and visual inspection of distributions and correlations.
- **04_Modelling_BaseModel.ipynb** Trains baseline models (Logistic Regression, LDA) and evaluates initial predictive performance.
- **04_Modelling_BinaryRating.ipynb** Implements binary classification (Investment Grade vs. Non-Investment Grade), including prediction and evaluation.
- **04_Modelling_Rating.ipynb** Performs multi-class rating prediction and evaluates five-class LDA and related models.

All notebooks are included with this project submission and are also publicly available in the GitHub repository linked below.

Full Project Repository: *Synthetic Data Generation and Prediction for Corporate Rating using Bayesian and MCMC Techniques*

18 References

- Full Project Repository: Synthetic Data Generation and Prediction for Corporate Rating using Bayesian and MCMC Techniques.
- MIT 6.0002 – Introduction to Computational Thinking and Data Science. MIT OpenCourseWare YouTube Playlist: Link.
- Unlocking the Potential of LSTM for Accurate Salary Prediction. PeerJ Computer Science. Link.
- MCMC-Based Credit Rating Aggregation Algorithm to Tackle Data Insufficiency. RePEc / Applied Research. Link.
- Joshua S. Speagle (2019). *A Conceptual Introduction to Markov Chain Monte Carlo Methods*. arXiv:1909.12313. Link.
- Metropolis–Hastings Algorithm. Wikipedia. Link.