Synthetic Data Generation and Prediction for Corporate rating using Bayesian and MCMC techniques

Ayushman Anupam (MDS202411)

Data Science
Chennai Mathematical Institute

Problem Statement

- Limited availability of real-world credit rating data due to confidentiality and sparse observations.
- Classical ML models require full retraining when new data arrives, making updates computationally expensive.
- Deterministic ML models cannot capture uncertainty, which is critical in financial risk modelling.
- Data often contains noise, missing values, and hidden latent factors, making classical models unstable.
- High risk of overfitting in traditional ML due to small datasets and high dimensionality.

Proposed Solution

- **Solution to Limited Data:** Synthetic credit rating data is generated using a Bayesian Metropolis-Hastings method.
- **Solution to Expensive Retraining in Classical ML:** Bayesian updating allows parameters to be refined incrementally using new observations, eliminating the need to retrain the entire model from scratch.
- **Solution to Lack of Uncertainty Modeling:** The MCMC-based Bayesian model provides posterior distributions for parameters and rating outcomes, capturing uncertainty instead of giving single-point predictions.
- **Solution to Noise, Missing Values, and Latent Factors:** Probabilistic modeling (likelihood + priors) naturally incorporates noise, handles missing information, and captures hidden structure more effectively than classical deterministic ML methods.
- **Solution to Overfitting in Small Datasets:** Bayesian priors act as regularization, and synthetic data augmentation combined with GMM + residual modeling reduces overfitting and stabilizes predictions.

Abstract

- **Industry Context:** Credit rating agencies rely heavily on limited, proprietary financial datasets and classical machine learning techniques that are often unable to keep pace with rapidly changing market conditions.
- **Existing Gaps:** Current models suffer from data scarcity, class imbalance, lack of uncertainty quantification, and the need for full retraining whenever new information becomes available, making the rating process slow and less adaptive.
- **My Contribution:** This work introduces a Bayesian–MCMC based framework that generates synthetic credit data, models uncertainty through posterior distributions, handles noise and latent structure probabilistically, and supports efficient incremental updating, thereby improving robustness and flexibility in credit rating prediction.

Introduction

I used Bayesian–MCMC framework to address limitations in real-world corporate credit rating data.

- **Synthetic datasets** are generated using GMM and Metropolis-Hastings MCMC to overcome confidentiality constraints and severe class imbalance.
- First, Bayesian modeling provides probabilistic predictions, allowing uncertainty quantification instead of single deterministic outputs.
- Then GMM clustering and residual correction to capture latent structure and stabilize rating accuracy.

This proposed system supports incremental updates, avoiding the need for complete retraining when new data becomes available.Results show improved robustness, reduced overfitting, and enhanced reliability in credit rating prediction.

Dataset

The dataset used in this project is sourced from Kaggle: Corporate Credit Rating with Financial Ratios.

The dataset consists of **7805 rows and 25 columns** with **no missing values**.
The columns include: Rating Agency, Corporation, Rating, Rating Date, CIK, Binary Rating, SIC Code, Sector, Ticker, 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, and Free Cash Flow Per Share.

Data Preparation

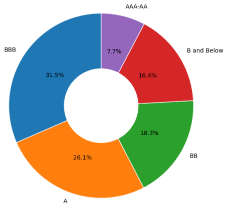
The initial preprocessing begins by dropping the columns **["Corporation", "Ticker", "CIK", "Rating Date"]**. Since the rating year is extracted separately, the full *Rating Date* column is no longer required.

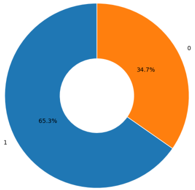
Categorical variables **["SIC Code", "Sector", "Rating_Year", "Rating Agency"]** are then transformed using one-hot encoding to make them suitable for modeling.

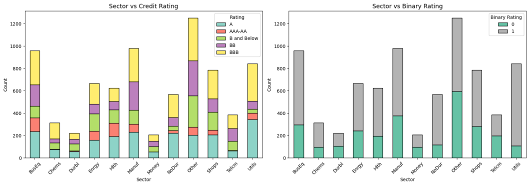
To simplify the multi-class rating problem, the original credit ratings are compressed into **five rating categories** using a mapping scheme, where related grades such as AAA, AA+, AA, and AA- are grouped into broader classes like *AAA-AA*, and lower ratings including B, CCC, CC, C, and D are merged into the *B and Below* class.

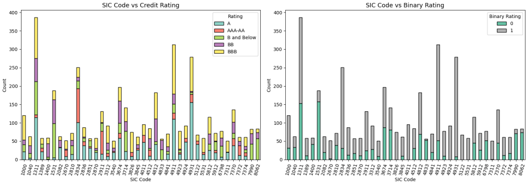
Two prediction targets are used: the original **Rating** and the **Binary Rating**. The new 5-class rating variable is encoded and stored as **Rating_Class_5** for downstream modeling.

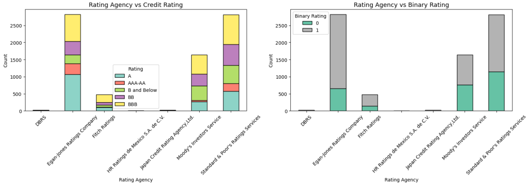
Data Analysis


Distribution of Credit Ratings


Distribution of Binary Ratings

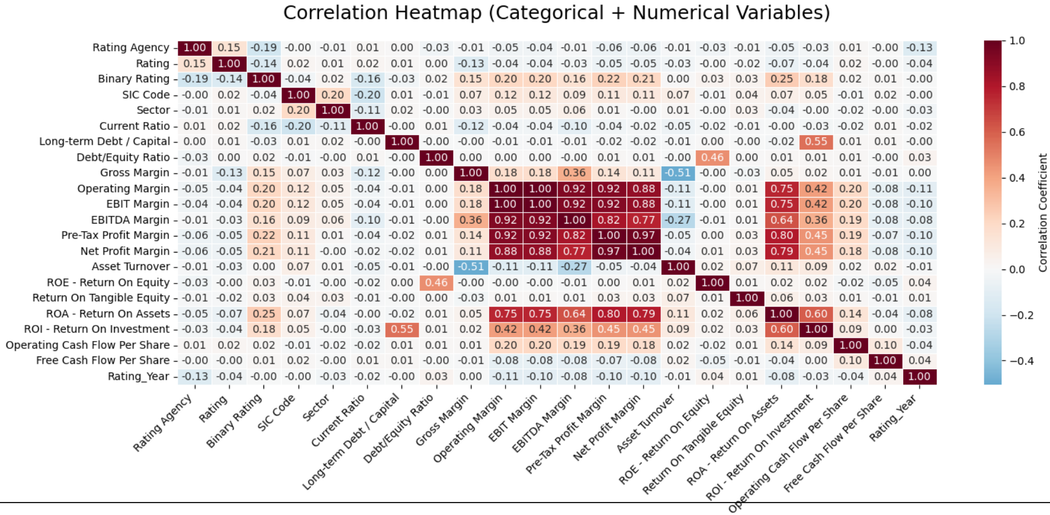

Sector vs Rating / Binary Rating


SIC Code vs Rating / Binary Rating


Rating Agency vs Rating / Binary Rating

Data Analysis - Continued

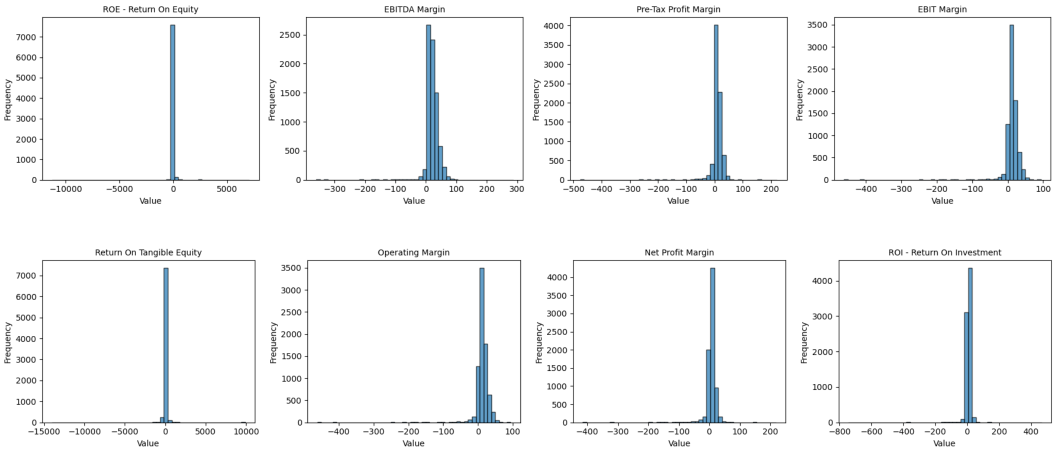
Correlation Heatmap



	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

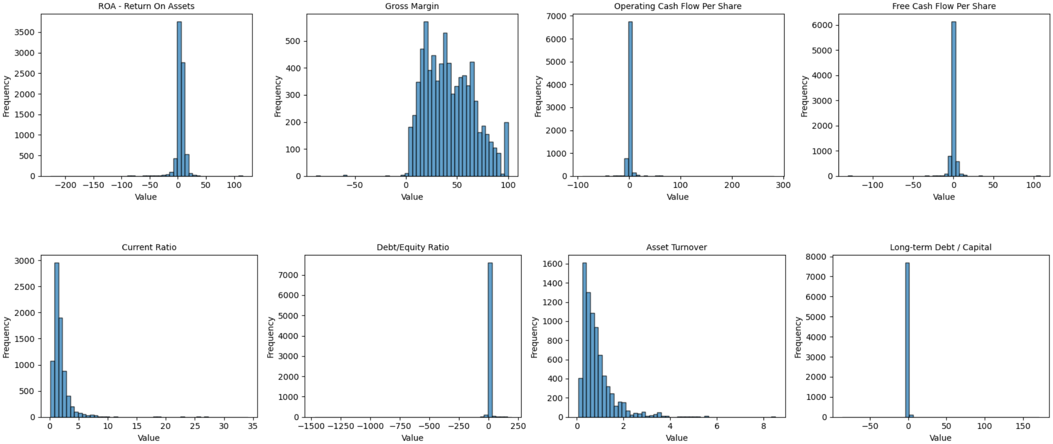
Data Analysis - Continued

Histogram of Numerical Data



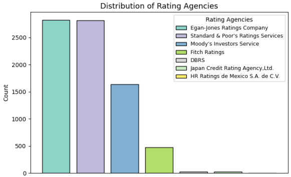
Data Analysis - Continued

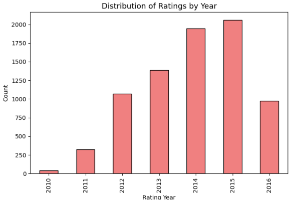
Histogram of Numerical Data - continued

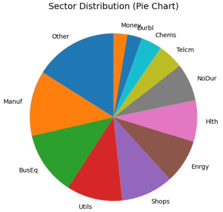


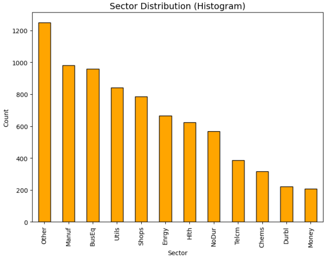
Data Analysis - Continued

Categorical Exploratory Data Analysis









Unique values in each categorical field:

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

Synthetic Data Generation

1. **GMM-Based Structural Modeling:** Gaussian Mixture Model is trained on data to component-wise numerical distributions, and categorical probability tables for one-hot and single-column categorical variables.
2. **Sampling Synthetic Observations:** Next, new samples are generated by sampling multivariate Gaussian for numerical features, and sampling categorical variables from component-specific probability distributions.
3. **Distributional Refinement:** Finally, quantile matching is used to align marginal distributions, repair one-hot encodings, and run an MCMC refinement step to correct variance and correlation structure in key financial ratios.
4. **Quality Evaluation:** Assess similarity between real and synthetic data using **Kullback-Leibler Divergence (KL)** and **Jensen–Shannon Similarity (JSS)** to ensure high-fidelity synthetic generation.

Synthetic Data Generation: GMM Modeling

I model the joint distribution of the dataset using a **Gaussian Mixture Model (GMM)** combined with component-wise categorical probabilities.

- Numeric features are normalized and the optimal number of mixture components is selected using BIC.
- For each GMM component, we estimate:
  - Mean and covariance for numeric variables,
  - Probability tables for one-hot and categorical features.
- This gives a flexible generative model capturing multimodality and mixed data types.

Synthetic Data Generation: Sampling and Alignment

1. **Sampling:**
    - Draw a mixture component based on GMM weights.
    - Sample numeric features from the component's multivariate normal.
    - Sample categorical variables from the component's probability tables.
2. **Distribution Alignment:**
    - Apply quantile-matching so each numeric column in the synthetic data closely follows the real marginal distribution.

This reduces distributional drift and ensures realistic synthetic observations.

Synthetic Data Generation: MCMC Refinement

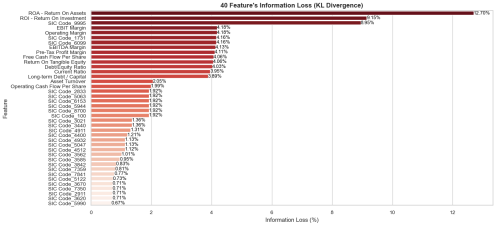
A lightweight MCMC procedure is applied to improve multivariate behaviour of data.

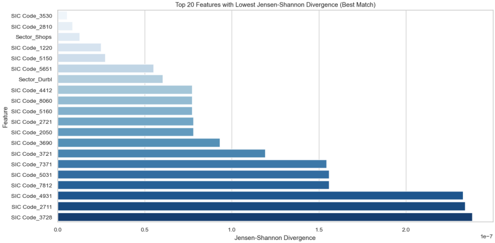
- Small random updates are proposed to selected numeric columns.
- A loss based on variance and correlation mismatch is computed.
- Proposals are accepted using a Metropolis rule.

The refinement step corrects covariance structure and improves realism in important financial ratios (ROE, ROI, ROA).

Finally, the refined synthetic data is merged with the real dataset for downstream modeling.

Synthetic Data Generation: Correctness





Across all major financial ratios (ROA, ROI, profit margins, leverage ratios), divergence values remain extremely low (e.g., KL $\approx$ 0.002–0.007), while similarity scores exceed **99%**. This demonstrates that the synthetic data preserves the statistical characteristics of the original dataset with high fidelity.

KL Divergence measures how much information is lost when the synthetic distribution approximates the real one. **Jensen–Shannon** Similarity measures how close the two distributions are in a symmetric and stable way.

Binary Rating Classification

1. Bayesian logistic regression model is used to model data: log-posterior = log-likelihood + Gaussian prior.
2. In next step, posterior samples of coefficients are drawn using using a Metropolis MCMC sampler.
3. then posterior probabilities and 95% credible intervals are calculated.
4. Finally, we refine model, by fitting a GMM on train residuals and apply penalty to refine test probabilities.

Binary Rating Classification - continued

Model (log-posterior):

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

(Here: Gaussian prior $\beta \sim \mathcal{N}(0, 10 \cdot I)$, $\sigma(\cdot) = \text{logistic}$)

Metropolis sampler (code):

- Start at $\beta^{(0)} = \mathbf{0}$, propose $\beta' = \beta + \mathcal{N}(0, \text{proposal\_std}^2 I)$.
- Compute $\Delta = \log p(\beta') - \log p(\beta)$, accept with probability $\min(1, e^{\Delta})$.
- Collect N samples, discard burn-in to obtain posterior samples $\{\beta^{(s)}\}$.

In my case: I did **30,000** iteration and burnin of **2000** for good acceptance.

Binary Rating Classification - continued

Posterior predictive probabilities:

$$p_i^{(s)} = \sigma(X_i \beta^{(s)}) \quad \text{for each posterior sample } s.$$

Point estimate: use posterior mean

$$\hat{p}_i = \frac{1}{S} \sum_{s=1}^{S} p_i^{(s)}.$$

95% credible interval: take empirical percentiles of $\{p_i^{(s)}\}_{s=1}^{S}$: **In my case:**

```
posterior_preds_test = expit(X_test @ posterior_samples.T)
lower = np.percentile(..., 2.5, axis=1); upper =
np.percentile(...,97.5,axis=1)
```

Binary Rating Classification - continued

Prediction refinement using GMM on Residuals `Residuals:` compute sample-wise residual matrix

$$r_i^{(s)} = y_i - p_i^{(s)} \quad \Rightarrow \quad \text{use aggregated residuals } r_i.$$

Train GMM on residuals:

- Fit a small GMM (e.g. 3 components) to all train residual values.
- For each test instance, compute posterior responsibilities γ_{ik} for GMM components using its residual samples.

Then, I use Penalty & refinement to refine prediction

Final output: table of y_i, $\hat{p}_i$, $\mathrm{CI}_{95\%,i}$, $\tilde{p}_i$, binary predictions and penalty — ready for evaluation.

Binary Rating Classification: Final Results

Original Data Evaluation

- 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)

Synthetic Data Evaluation

- 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)

Multiclass Rating Classification

Pipeline overview (brief):

1. **Model:** Bayesian multinomial logistic regression with a weak Gaussian prior to model class probabilities.
2. **Sampling:** Use a Metropolis–Hastings MCMC sampler to draw posterior samples of the coefficient matrix B.
3. **Posterior Prediction:** For each test sample, compute softmax probabilities for every posterior draw and take the posterior mean as the predicted probability. I later use it, to compute **95% credible interval** for each class.

Multiclass Rating Classification - continued

We model the credit rating as a **multiclass logistic regression** using a Bayesian framework.

- For K classes, we estimate $(K-1)$ parameter vectors; the last class acts as the reference category.
- Linear scores are mapped to class probabilities using the softmax function.
- A weak Gaussian prior $B \sim \mathcal{N}(0, 10)$ is placed on all coefficients to regularize the model.

The task is to compute the posterior distribution of all coefficients by MCMC.

Multiclass Rating Classification - continued

We obtain posterior samples of the coefficient matrix B using a **Metropolis–Hastings sampler**.

- Start with $B^{(0)} = 0$.
- Propose $B' = B + \mathcal{N}(0, \sigma^2)$ for all coefficients.
- Accept the proposal with probability $\min\{1, \exp(\log p(B') - \log p(B))\}$.
- After burn-in, the chain provides posterior draws $\{B^{(s)}\}_{s=1}^{S}$.

These samples represent the uncertainty in rating-class boundaries.

Multiclass Rating Classification - continued

For each test observation:

- For every posterior sample $B^{(s)}$, compute class probabilities using the softmax function.
- This gives a posterior distribution of predicted probabilities for every class.
- The final probability is the posterior mean: $\hat{p}_{i,c} = \frac{1}{S} \sum_s p_{i,c}^{(s)}$.
- The **95% credible interval** is obtained from the 2.5th and 97.5th percentiles of the posterior predictive samples.

Thus every prediction comes with a full uncertainty quantification.

Multiclass Rating Classification – 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)

Accuracy: 0.5452

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)

Accuracy: 0.4795

Base Model: Binary Rating Classification

Model Used: Logistic Regression

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

Base Model: Multiclass Rating Classification

Model Used: Linear Discriminant Analysis (LDA)

Original Data

- Accuracy: 0.5586
- Precision (macro): 0.5394
- Recall (macro): 0.5318
- F1 Score (macro): 0.5327

Sample Predictions:

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

Synthetic Data

Evaluation:

- Accuracy: 0.485
- Precision (macro): 0.4684
- Recall (macro): 0.4906
- F1 Score (macro): 0.4752

Sample Predictions:

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

Result

Model Comparison with Base Models

Original Data Comparison

Binary Rating (4 metrics)

Metric	Base Model	my Model
Accuracy	0.8430	**0.8949**
Precision	0.8589	**0.9388**
Recall	0.9143	**0.9008**
F1 Score	0.8857	**0.9194**

Multiclass Rating (1 metric)

Metric	Base Model	my Model
Accuracy	0.5586	**0.5452**

Synthetic Data Comparison

Binary Rating (4 metrics)

Metric	Base Model	my Model
Accuracy	0.7809	**0.8077**
Precision	0.7955	**0.8971**
Recall	0.8959	**0.7981**
F1 Score	0.8427	**0.8447**

Multiclass Rating (1 metric)

Metric	Base Model	my Model
Accuracy	0.4850	**0.4795**

Results Summary

- **High-Quality synthetic data** was generated using a hybrid GMM + quantile-matching + MCMC refinement pipeline. KL divergence remained in the range **0.002–0.007** and Jensen–Shannon similarity exceeded **99%**, confirming strong distributional alignment with real data.
- **Binary Rating Model (Bayesian Logistic Regression + Residual GMM Correction)** showed clear improvement over the base Logistic Regression:
  - **Original Data:** Accuracy improved from 0.8430 → **0.8949**
  - **Synthetic Data:** Accuracy improved from 0.7809 → **0.8077**
  - **95% credible intervals**, enabling uncertainty-aware decisions.
- **Multiclass Bayesian Rating Model** produced full probability distributions and credible intervals for all five rating classes. Its accuracy was lower than the binary model—**mainly because the dataset contains ratings from multiple agencies with inconsistent grading scales**. If ratings were sourced from a **single agency with uniform rating criteria**, the multiclass model would likely perform significantly better.

Conclusion

- **Bayesian MCMC models offer a more reliable framework** for corporate credit rating prediction, especially when data is limited, noisy, or expensive to obtain.
- The proposed **synthetic data generation pipeline** solves the common industry problem of data scarcity by producing statistically realistic samples.
- Unlike classical ML models that must be retrained from scratch,Bayesian updating allows **incremental learning**, making it more adaptive to new data.
- **posterior intervals, distributional correction via GMM**, and **uncertainty estimation** makes predictions more transparent and defensible for risk-sensitive applications.
- The model demonstrated strong improvements in binary rating prediction and competitive multiclass performance, showing that **Bayesian methods can outperform or match traditional models while providing deeper insight.**

Overall, the system provides a practical, uncertainty-aware, and data-efficient approach that can be applied in **banking, credit risk rating agencies, NBFCs, investment research, and financial stress-testing frameworks.**

Problem in Industry, I tried to solve

Financial institutions and rating agencies face several practical challenges in real-world credit risk modeling:

- **Limited and confidential credit rating data** restricts the development and validation of robust statistical models.
- **Frequent arrival of new financial statements** requires continuous model updates, making classical ML retraining computationally expensive and slow.
- **Deterministic predictions fail to quantify uncertainty**, which is essential for risk-sensitive decisions such as lending, portfolio allocation, and regulatory reporting.
- **Noisy, missing, or inconsistent financial ratios** make traditional models unstable and prone to overfitting.
- **Need for explainable and probabilistic forecasting** is growing, especially under Basel/IFRS regulations and stress-testing frameworks.

my Bayesian and synthetic data generation framework directly addresses these industrial pain points.

Industrial Application

The proposed Bayesian–MCMC credit rating framework directly addresses key industry challenges such as limited labeled data, costly retraining of classical ML models, and the need for uncertainty-aware decision making.

Where this model can be applied:

- **Banks and NBFCs:** Improve internal credit scoring, assess borrower risk, and quantify uncertainty in rating transitions.
- **Credit Rating Agencies:** Generate realistic synthetic datasets for model development, stress testing, and regulatory audits without exposing confidential data.
- **Investment Firms & Asset Managers:** Use uncertainty-aware credit predictions to adjust portfolios, rebalance exposures, and identify high-risk issuers.
- **FinTech and Alternative Lending Platforms:** Continuously update models using Bayesian incremental learning, avoiding full retraining as new borrower data arrives.

References

- **Full Project Repository:** GitHub Repository `github.com/AyushmanGHub/Synthetic-Data-Generation-and-Prediction-for-Corporate-rating-using-Bayesian-and-MCMC-techniques`
- **MIT 6.0002 – Introduction to Computational Thinking and Data Science:** YouTube playlist by MIT OpenCourseWare
- **Unlocking the Potential of LSTM for Accurate Salary Prediction:** (MLE, Jeffreys prior, and advanced risk functions). PeerJ Computer Science
- **MCMC-Based Credit Rating Aggregation Algorithm to Tackle Data Insufficiency:** RePEc / Applied Research
- **A Conceptual Introduction to Markov Chain Monte Carlo Methods** Joshua S. Speagle Harvard Cambridge
- **Metropolis–Hastings Algorithm:** Wikipedia