Who are the Loan Defaulters? - An MCMC: Metropolis Hastings Approach

CSCI-5822 Final Project

Josh-Myers Dean, Samuel Kwon, Chandra Kanth Nagesh

Department of Computer Science
University of Colorado Boulder, CO - 80302

05/01/2023

University of Colorado
Boulder

Outline

Introduction

Background Information

- **Defaulting** is the failure to repay interest or principal components on a loan or any borrowed finance.

Objectives

- How can we apply a probabilistic model to predict loan defaulting?
- Can we effectively estimate the parameters for logistics regressions using MCMC for our task?
- How accurate and flexible is the model in predicting loan defaulting behavior given consumer financial metrics?

Why is This Useful?

- An automated approach to assess loan default behavior, enabling faster risk calculation and improved decision-making processes.
- By leveraging a Bayesian approach, we can account for more uncertainty in our decision-making process compared to a frequentist approach.
- Caveat: This probably shouldn't be used in the real world (i.e., discriminatory practices).

Outline

Exploratory Data Analysis

- Loan Default Dataset from Kaggle.
- Dataset contains 70 features with a mixture of continuous and discrete features.
- Random variables include metrics about the approved loan and metrics about the applicant.
- Contains around 148,671 rows of data – 70/30 split for training and testing

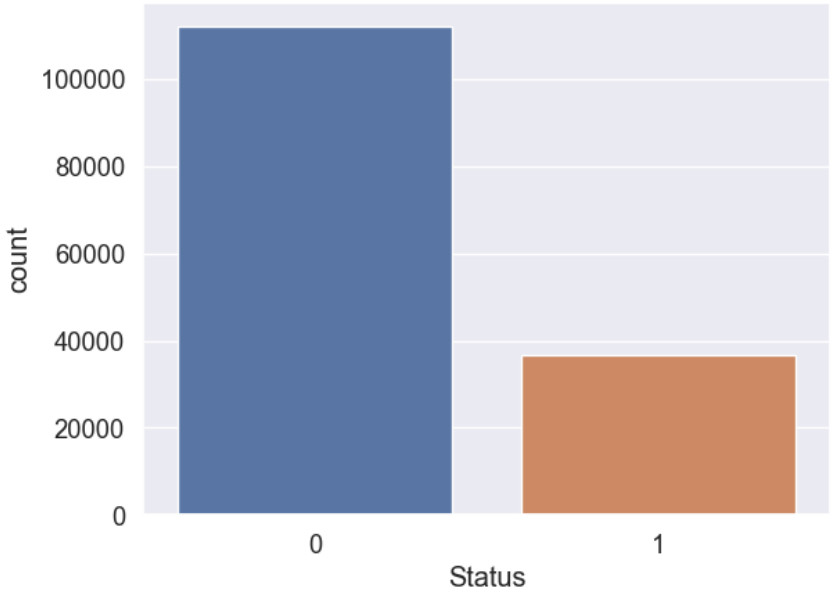


Figure: Frequency of Default Status In Our Data

Exploratory Data Analysis

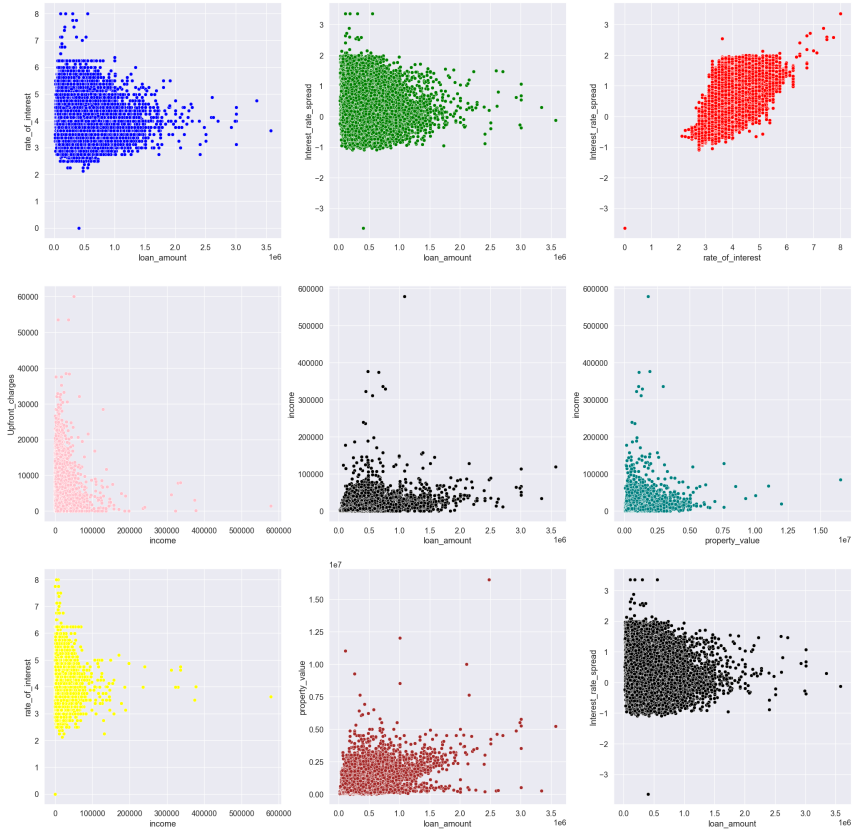


Figure: Scatter plot of Key Features

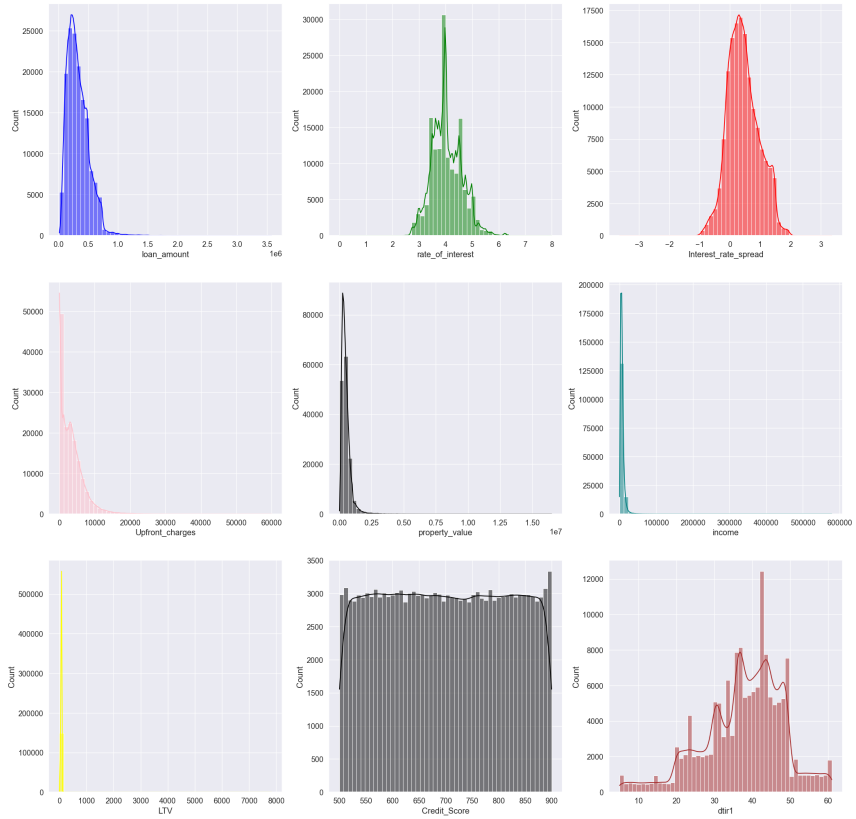


Figure: Histogram of Key Features

Outline

Prinicipal Component Analysis (PCA)

- We perform informed PCA by looking into the **Scree plots**.
- The point of inflection is around 30 components, which explains 80% of the data.
- We then perform PCA and pick the 30 features based on their explained variance ratios, to form the dataset.
- Features: *Income, Term, Loan Amount...*

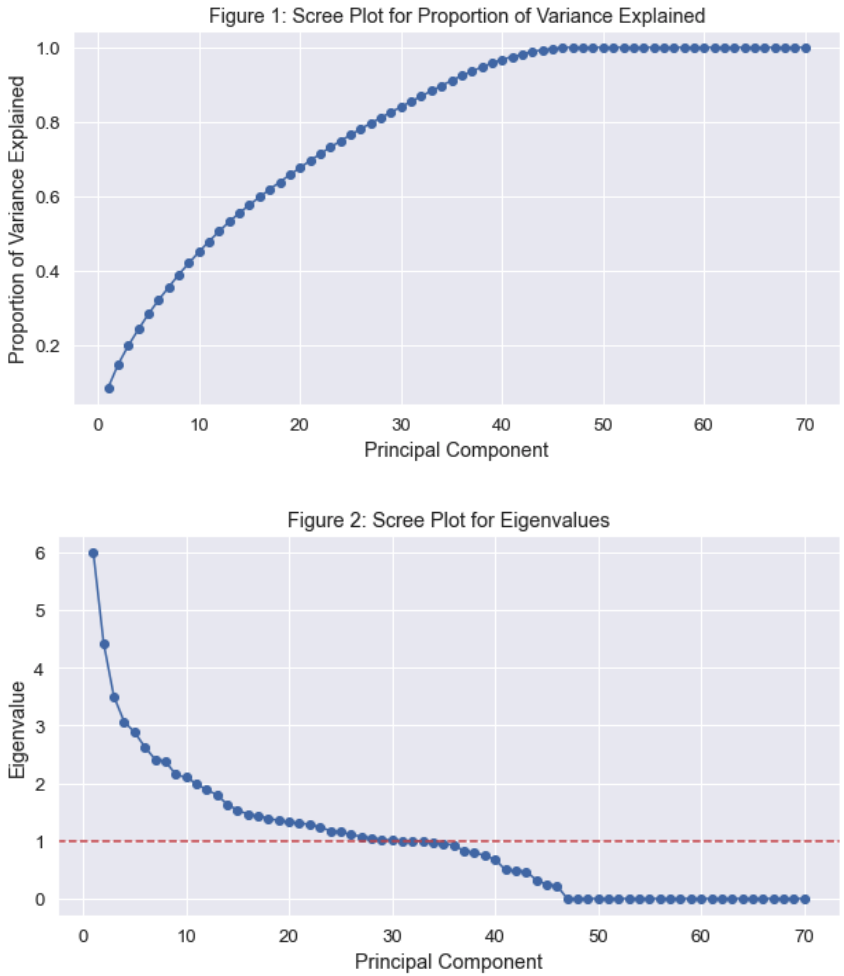


Figure: Explained variance per Principal Compnents

Outline

Methods

Logistic Regression

We estimate the β parameters in logistic regression using MH-MCMC.

$$\underline{\text{Joint Prior Distribution:}} \quad p(\beta_i) = \frac{1}{\sigma_{\beta_i}\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{\beta_i - \mu_{\beta_i}}{\sigma_{\beta_i}}\right)^2} \equiv \beta_i \sim \mathcal{N}(0,1) \quad (1)$$

$$\underline{\text{Model:}} \quad y_i = L(P) = \frac{1}{1+e^{-P}}, \; P = \beta_0 + \sum_{i=1}^{30} \beta_i x_i \quad (2)$$

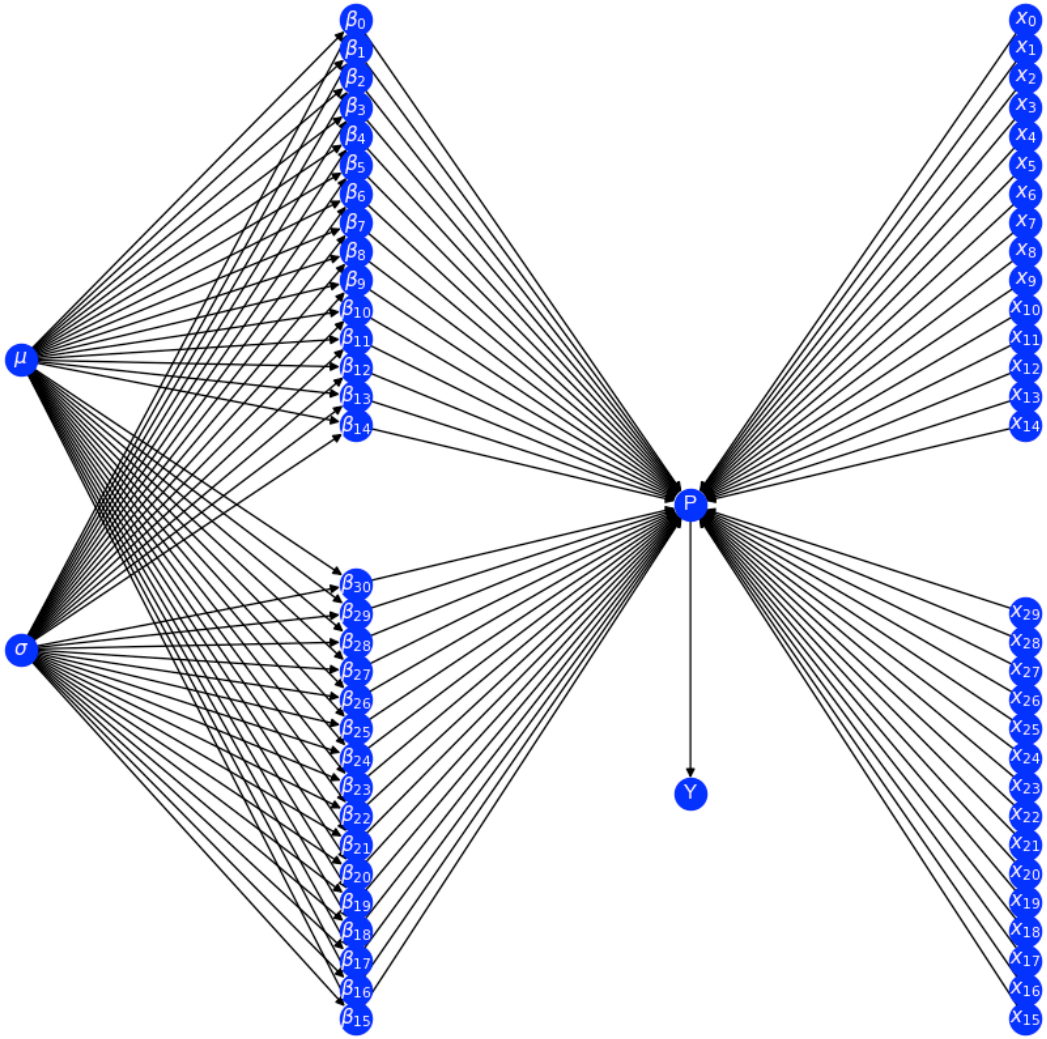
Markov Chain Monte Carlo Method: Metropolis Hastings Algorithm (PyMC3)

Sampling : 10,000
Burn In Period : 500
Chains : 4

Outline

Results

Belief Network



Results

Moralized Graph

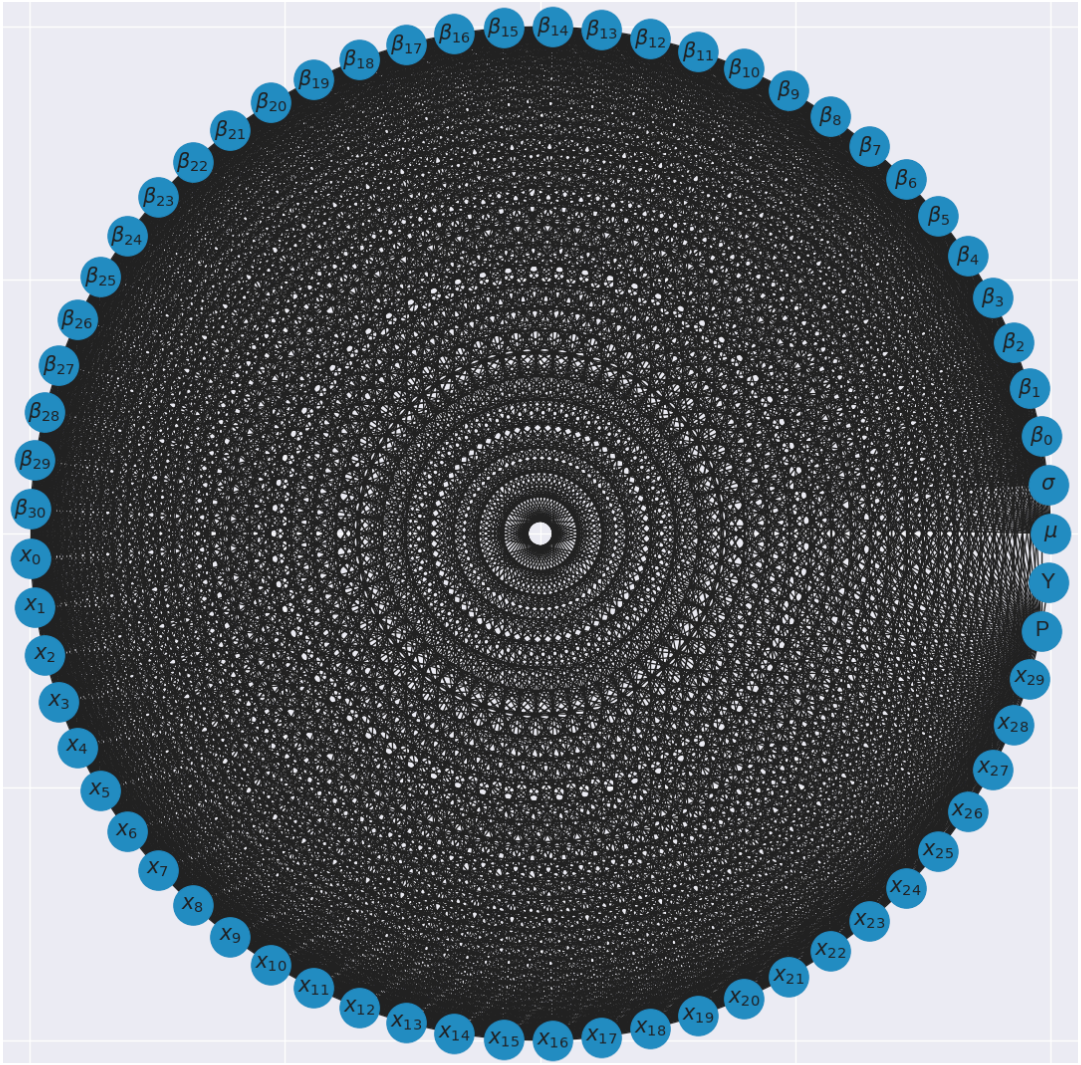
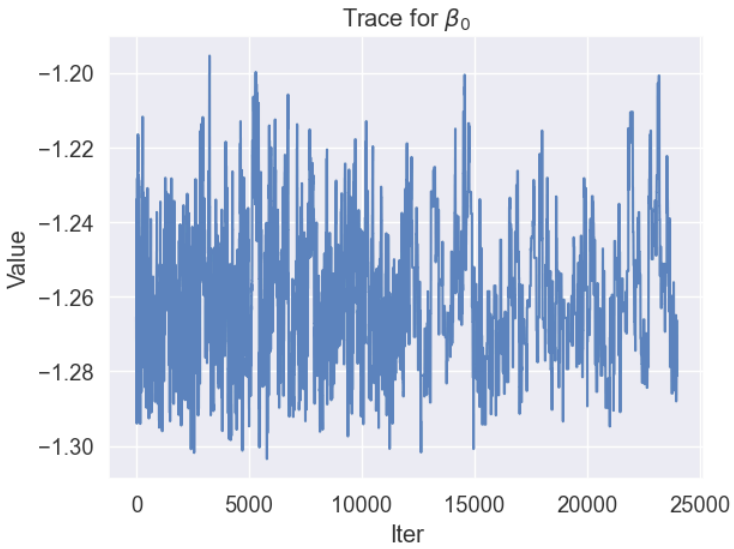
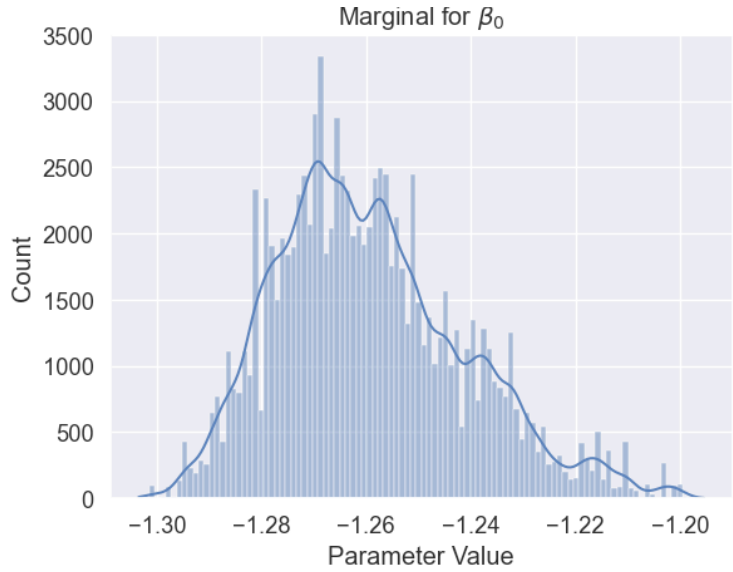
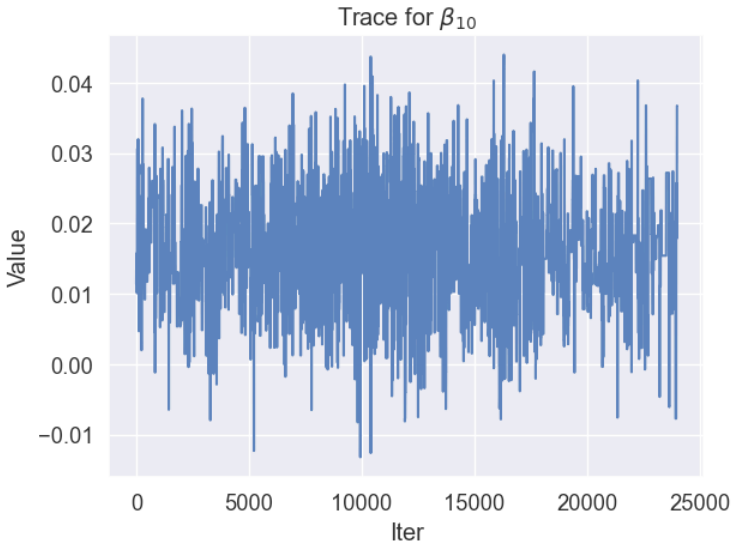


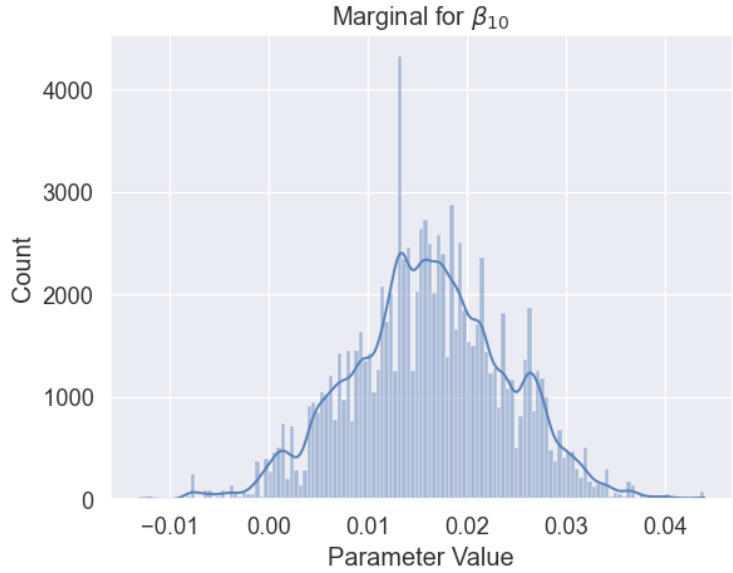
Figure: Moralized Network

Results

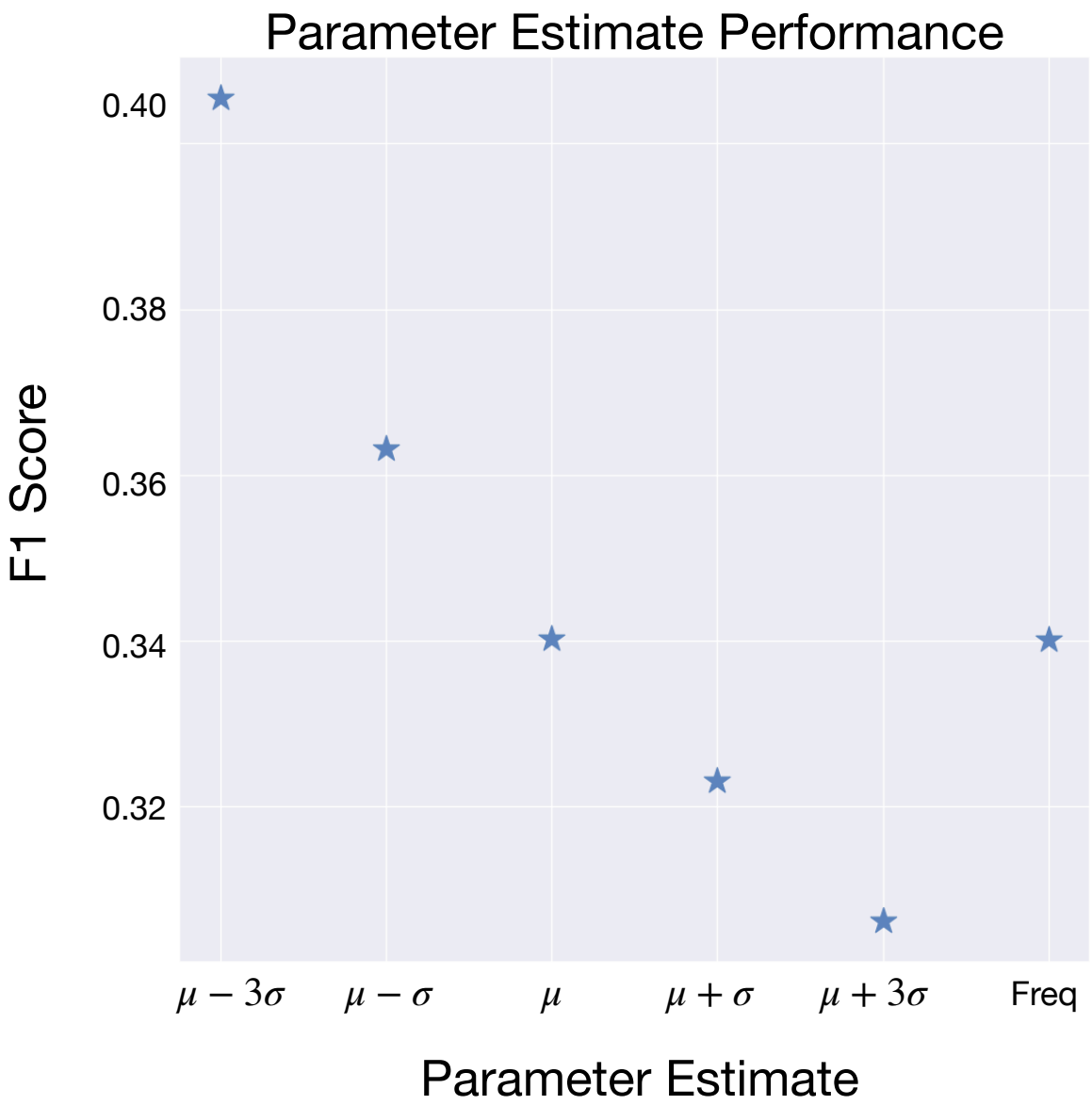

Trace for β_0
Value
−1.20
−1.22
−1.24
−1.26
−1.28
−1.30
0
5000
10000
15000
20000
25000
Iter


Marginal for β_0
Count
3500
3000
2500
2000
1500
1000
500
0
−1.30
−1.28
−1.26
−1.24
−1.22
−1.20
Parameter Value


Trace for β_{10}
Value
0.04
0.03
0.02
0.01
0.00
−0.01
0
5000
10000
15000
20000
25000
Iter


Marginal for β_{10}
Count
4000
3000
2000
1000
0
−0.01
0.00
0.01
0.02
0.03
0.04
Parameter Value

Results


Parameter Estimate Performance
F1 Score
0.40
0.38
0.36
0.34
0.32
$\mu - 3\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 3\sigma$
Freq
Parameter Estimate

Results

Parameter Estimate	Label	Precision	Recall	F1 Score	Support
$\mu - 3\sigma$	0	**0.80**	0.95	0.87	33532
	1	0.65	**0.29**	**0.41**	11068
$\mu - \sigma$	0	0.79	0.96	0.87	33532
	1	0.69	0.25	0.36	11068
μ	0	0.79	0.97	0.87	33532
	1	0.70	0.22	0.34	11068
$\mu + \sigma$	0	0.79	0.97	0.87	33532
	1	0.70	0.21	0.32	11068
$\mu + 3\sigma$	0	0.79	0.97	0.87	33532
	1	0.70	0.20	0.31	11068
Frequentist	0	0.79	0.97	0.87	33532
	1	0.70	0.22	0.34	11068

Table: Results on test set using varying point estimates for each parameter. 0: Didn't Default, 1: Defaulted.

Outline

Discussion and Conclusion

- Frequentist approach (Sklearn) matches the Bayesian approach when considering the mean.
- Bayesian approach allows to tune our results with different point estimates from our parameter distributions, better matching the true outcomes.
- Model predicts label 0 (Not Defaulted) with higher precision than label 1 (Defaulted), likely due to data imbalance.

Outline

References

- https://www.kaggle.com/datasets/yasserh/loan-default-dataset
- PyMC3
- NetworkX
- https://github.com/joshmyersdean/CSCI5822-FinalProject