Bayesian Inference Analysis of UPSC Selection Probability

1. Objective

We aim to model the probability of clearing the UPSC exam using Bayesian Logistic Regression.

2. Synthetic Data Generation

Since real data is unavailable, we generated synthetic data (data for 2000 candidates) representing realistic candidate behaviour. The following features were considered.

- Study hours per week (mean = 50, Standard deviation = 5)
- Mock test (Between 0 to 1, 0.6 means on an average 60% marks in mock tests)
- Social media usage (log-normal distributed)
- Relationship breakup status (0 or 1)

True Model (Hidden Relationship)

We assumed the true log-odds model:

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot (\text{Study}) + \beta_2 \cdot (\text{Mock}) + \beta_3 \cdot (\text{Social}) + \beta_4 \cdot (\text{Breakup})$$

Where:

- $\beta_{study} = 0.08$
- $\beta_{mock} = 3.0$
- $\beta_{social} = -0.25$
- $\beta_{breakup} = 0.4$

The intercept β_0 was computed numerically to ensure that:

$$\text{Average Probability} = \text{Based on the actual selection rate from dataset}$$

3. Meaning of Coefficients (Odds Interpretation)

Logistic regression models **odds** instead of probability.
Odds are defined as:

$$\text{Odds} = \frac{p}{1-p}$$

where p is probability of success.
If $\beta = 0.08$ for study:

$$e^{0.08} = 1.083$$

Meaning: Each additional study hour increases the odds of selection by 8.3%. Similarly:

- Mock performance strongly increases odds.
- Social media reduces odds.
- Breakup gives small positive psychological effect.

Every factor multiplies your "chances ratio" rather than directly adding probability.

4. Bayesian Assumptions

Before seeing data, we assumed:

$$\beta_i \sim \mathcal{N}(0, \sigma^2)$$

This means:

- Initially, we believe all features are equally important.
- We allow data to update this belief.

Bayes' rule updates prior belief using observed data to produce the posterior distribution.

5. Feature standardization

Continuous features were standardized before modeling. We put all factors on the same scale so none dominates just because of units.

6. Interpretation of Results

Posterior Distributions

Each graph shows:

- Dashed curve = prior belief
- Histogram = posterior belief after seeing data
- Vertical lines = prior mean and posterior mean

What We Observe

Table 1: Prior vs Posterior Estimates of Logistic Regression Coefficients (N = 2000)

Parameter	**Prior Mean**	**Posterior Mean**	**Posterior SD**	**Technical Interpretation**	**Layman Interpretation**
Intercept (β_0)	0	-7.965	1.028	Baseline log-odds = -7.965 → Odds ratio = exp(-7.965) $\sim$ 0.00035. Very low baseline probability when all predictors = 0. Data strongly shifted belief from neutral prior.	If none of the factors are present, the event is extremely unlikely.
Study Hours (β_{study})	0	0.573	0.480	Odds ratio = exp(0.573) $\sim$ 1.77. Studying increases odds by $\sim$ 77%. Moderate uncertainty but effect likely positive.	Studying makes the event noticeably more likely.
Mock Tests (β_{mock})	0	0.179	0.490	Odds ratio = exp(0.179) $\sim$ 1.20. About 20% increase in odds. Effect small and uncertain.	Mock tests slightly increase the chances, but effect is weak.
Social Media Usage (β_{social})	0	-1.831	0.873	Odds ratio = exp(-1.831) $\sim$ 0.16. About 84% reduction in odds. Strong negative effect.	Social activity greatly reduces the chance of the event.
Breakup Stress ($\beta_{breakup}$)	0	0.618	0.958	Odds ratio = exp(0.618) $\sim$ 1.86. Nearly doubles odds, but high uncertainty (large SD).	Breakup may increase the chances, but we are less certain about this effect.

- Without any of these factors, the event is extremely unlikely.
- Studying increases the chances noticeably, mock tests help only slightly, and social activity significantly reduces the chances.
- A breakup may increase the chances considerably, but we are less certain about this effect compared to studying or social activity.

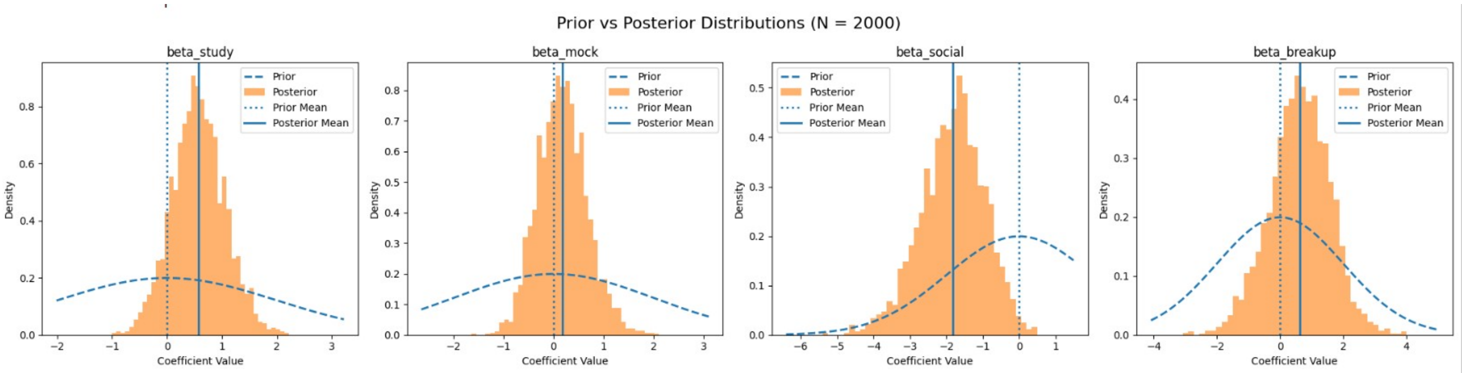
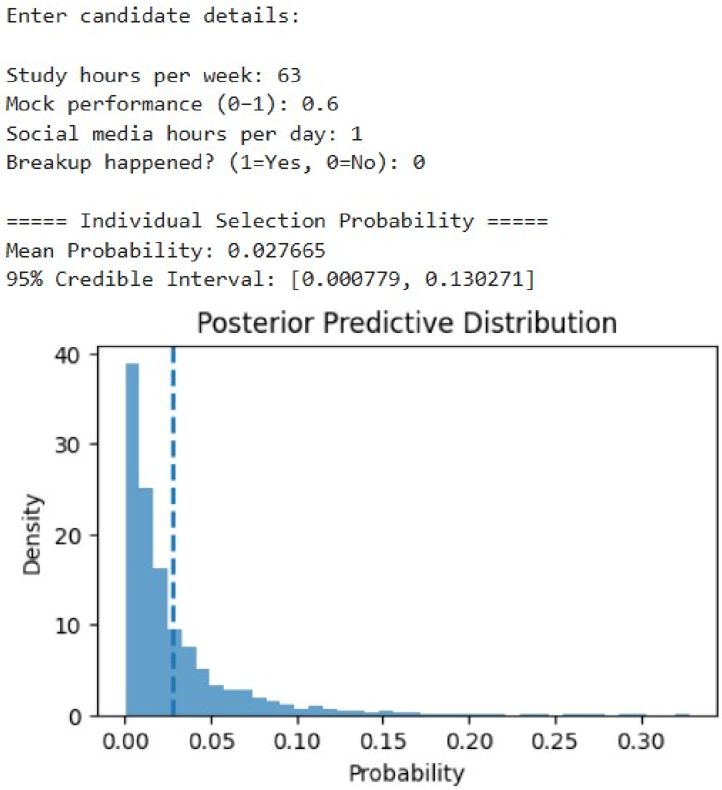


Figure 1: Prior and posterior distribution plots

7. Individual candidate probability distribution

Individual candidates can enter their details and check their probability of getting selected in UPSC CSE examination.
Link: https://colab.research.google.com/drive/1ziPGhvogxVUCU_HidxgLDgCBl9q4Em5g?usp=sharing
For example: Figure 2 shows the demo of probability predictor.



```
Enter candidate details:

Study hours per week: 63
Mock performance (0-1): 0.6
Social media hours per day: 1
Breakup happened? (1=Yes, 0=No): 0

===== Individual Selection Probability =====
Mean Probability: 0.027665
95% Credible Interval: [0.000779, 0.130271]
```

Figure 2: Probability predictor