Bayesian Data Analysis course - Project work at Aalto

Thilini Panagoda, Vivian Phan, Alexandr Pimikov [1]

1 Introduction

The assessment and control of default risk is a critical function in the ever-changing financial services industry, especially in the context of credit card lending. This study aims to identify the complex variables affecting an individual's creditworthiness within the particular setting of Taiwanese consumers' default payments.

This report presents a comparative study of two statistical models designed to predict credit default. The first is a simple logistic regression model, and the second is a more complex hierarchical model that takes into account the educational background of credit applicants. Our goal is to determine the impact of borrowers' education on their likelihood of making timely payments. We particularly focus on the practical relevance of including educational background as a hierarchical element in the model and whether this additional complexity translates into better predictive accuracy compared to the simpler logistic regression approach.

2 Description of the data and the analysis

We analyze the data on default of credit card clients taken from [1]. These data has been studied in [2, 3]. The previous studies used machine learning methods, while we used Bayesian data analysis methods.

The data include 30000 entries on 23 explanatory variables:

- Amount of the given credit (LIMIT_BAL)
- Clients data (SEX, EDUCATION, MARRIAGE, AGE)
- History of payment (PAY_0 PAY_2 PAY_3 PAY_4 PAY_5 PAY_6)
- Amount of bill statement (BILL_AMT1 BILL_AMT2 BILL_AMT3 BILL_AMT4 BILL_AMT5 BILL_AMT6)
- Amount of previous payment (PAY_AMT1 PAY_AMT2 PAY_AMT3 PAY_AMT4 PAY_AMT5 PAY_AMT6)
- default payment (Yes = 1, No = 0)

The correlation matrix plot provides a visual representation of the relationships between different features in the dataset. The attribute of interest, 'default', which represents the outcome we aim to predict, shows varying degrees of correlation with different features. Notably, it exhibits a significant correlation with the 'PAY' attributes, which represent the repayment status in past periods. Among these, 'PAY_0' and 'PAY_2' emerge as the most strongly correlated with 'default', suggesting their higher predictive power for the outcome variable. Consequently, these two features have been selected for further analysis.

[1] Email: pimikov@gmail.com

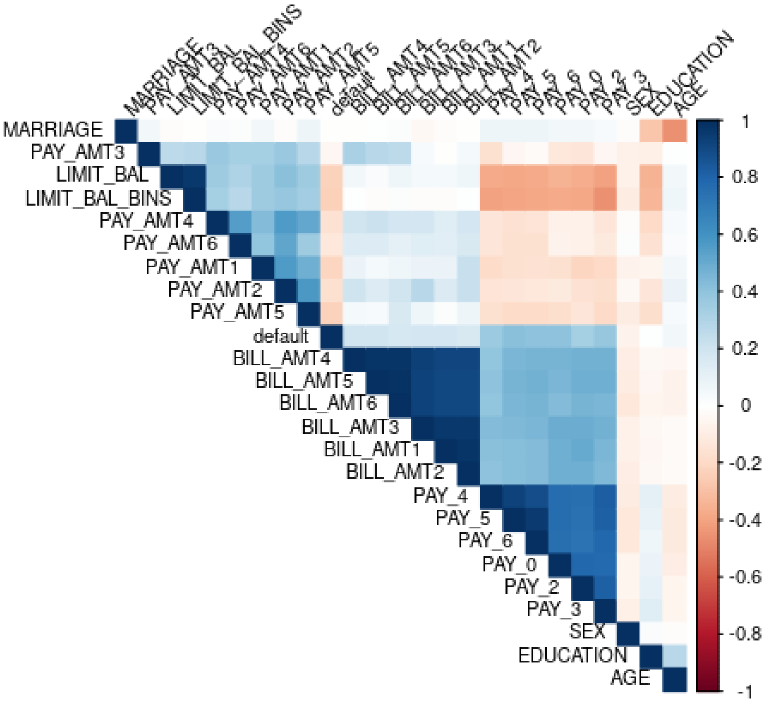


Figure 1: Correlation Matrix Plot

In addition to the payment history, background attributes like ’SEX’, ’EDUCATION’, and ’AGE’ are considered. ’EDUCATION’ stands out as the attribute most correlated with both ’PAY_0’ and ’PAY_2’. This insight has guided the decision to stratify the hierarchical model by levels of education, hypothesizing that educational background may interact with repayment behavior in predicting default.

Our analysis focuses on a small sampled subset of the data, ensuring representation across different educational backgrounds (1 for graduate school, 2 for university, 3 for high school, 4 for others) and default outcomes (0 for no default, 1 for default). For each category within education and default status, we sampled 40 data points, culminating in a dataset of 240 points. This approach allows for a balanced view across the spectrum of educational attainment and repayment performance.

The models were initially fitted using this compact dataset to capture the nuances of the relationship between education level, recent payment history, and default likelihood. Post model fitting, we proceeded to validate our findings on the full dataset, comprising 30,000 entries. This step is crucial to ascertain the model’s robustness and its generalizability to a broader population.

3 Description of two models

We use two models: the pooled model and hierarchical model. Our pooled model has the following form

$$\begin{aligned} &\mu = \text{logistic}(\beta_0 + \beta_1 * \text{PAY\_0} + \beta_2 * \text{PAY\_2})\,, \\ &d \sim \text{Bernoulli}(\mu) \end{aligned}$$

where logistic function

$$\text{logistic}(x) = \frac{1}{1 + exp(-x)},$$

and d stays for default. Hierarchical model has one extra layer for intercept grouped by education $i = 1, 2, 3$:

$$\begin{aligned}
&d_i \sim \text{Bernoulli}(\mu_i)\\
&\mu_i = \text{logistic}(\beta_0 + \beta_1 * \text{PAY\_0} + \beta_2 * \text{PAY\_2} + \alpha_i),\\
&\alpha_i \sim \mathcal{N}(a_i, \sigma).
\end{aligned}$$

4 Priors selection

Table 1: Dataset descriptions

Dataset	**Nof. Entries**	**Nof. 0 entries**	**Nof. 1 entries**
Short Data	240 entries	120 (50%)	120 (50%)
Original Data	30000 entries	23364 (77.88%)	6636 (22.12%)

Two distinct types of priors have been employed: a non-informative prior, $\mathcal{N}(0, 100)$, and a weakly informative prior, $\mathcal{N}(\log(\frac{0.7}{0.3}), 1)$. These priors are instrumental in guiding the Bayesian inference process. Below is a detailed description of each prior, along with their justifications and mathematical formulations:

4.1 Non-Informative Prior $\mathcal{N}(0, 100)$

This prior is a normal distribution with a mean (μ) of 0 and a substantial standard deviation (σ) of 100. It represents a non-informative or weakly informative stance, implying limited prior knowledge about the parameter's value.

The mean of 0 indicates no initial preference for positive or negative values of the parameter, suggesting a neutral starting point. The large standard deviation of 100 reflects significant uncertainty regarding the parameter, essentially allowing the data to primarily influence the posterior distribution.

The probability density function (PDF) for this prior is given by:

$$\mathcal{N}(0, 100) : P(\theta) = \frac{1}{\sqrt{2\pi \times 100^2}} \exp\left(-\frac{(\theta - 0)^2}{2 \times 100^2}\right)$$

where θ represents the parameter being estimated.

$$\beta_i \sim \mathcal{N}(0, 100).$$

where $i = 0, 1, 2$.

For the hierarchical model we have additional parameters, for which following noninformative priors used to pertain to group-level effects (random effects):

$$a_i \sim \text{Exponential}(0.02)$$
$$\sigma \sim \text{Exponential}(0.02)$$

4.2 Weakly Informative Prior: $\mathcal{N}(\log(\frac{0.7}{0.3}), 1)$

Regarding the datasets distribution in Table 1, the short dataset comprises 240 entries, evenly split between the two classes of 'default' (50% for class 0 and 50% for class 1), providing a balanced case scenario for model training. In contrast, the full dataset contains 30,000 entries with a class imbalance, where class 0 (non-default) constitutes the majority with 77.88%, and class 1 (default) is the minority with 22.12%. The choice of the prior distribution for the full dataset as $\mathcal{N}(\log(\frac{0.7}{0.3}), 1)$ is reflective of this observed class distribution. The log transformation within the normal distribution's mean parameter is an attempt to adjust for the skewness towards the majority class, potentially improving the model's ability to predict the less frequent event of 'default'.

This prior is a normal distribution centered around the natural logarithm of the odds ratio derived from a known class distribution (70% vs. 30%). It has a standard deviation of 1, indicating moderate confidence in this prior belief.

The PDF for this prior is:

$$\mathcal{N}(log(\frac{0.7}{0.3}), 1) : P(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\theta - \log(\frac{0.7}{0.3}))^2}{2}\right)$$

$$\beta_i \sim \mathcal{N}(\frac{0.7}{0.3}, 1)\,,$$
$$\beta_0 \sim \mathcal{N}(0, 10)\,.$$

where $i = 1, 2$.

The performance of both priors are quite similar (report further below in Sensitivity analysis). Therefore, the report will concentrate only on the prior $\mathcal{N}(0, 100)$

5 MCMC inference

In the pooled model, we use 4 chains in total, each with 2000 iterations and 1000 warm-up iterations.

In the hierarchical model, in order to to reduce the value of $\hat{R}$ from 1.02 for case of hierarchical model we increased the number of iterations to 5000 and warmup up to 2500.

5.1 Rhat and ESS value for convergence diagnostic

Convergence diagnostic is summarized in table 2 for the pooled model and table 3 for the hierarchical model.

	variable	mean	median	sd	mad	q5	q95	rhat	ess_bulk	ess_tail
1	b_Intercept	-0.21	-0.21	0.15	0.15	-0.46	0.03	1.00	2592.41	2472.17
2	b_PAY_0	0.52	0.52	0.15	0.16	0.28	0.78	1.00	2058.43	2543.86
3	b_PAY_2	0.16	0.16	0.13	0.13	-0.05	0.38	1.00	2074.70	2361.45
4	lprior	-16.57	-16.57	0.00	0.00	-16.57	-16.57	1.00	2417.79	
5	lp__	-158.24	-157.93	1.21	0.98	-160.77	-156.90	1.00	1876.25	2311.99

Table 2: Pooled model

	variable	mean	median	sd	mad	q5	q95	rhat	ess_bulk	ess_tail
1	b_I	-0.42	-0.39	0.74	0.36	-1.50	0.63	1.01	505.62	232.77
2	b_PAY_0	0.22	0.23	0.17	0.16	-0.05	0.51	1.00	1088.72	281.44
3	b_PAY_2	0.59	0.59	0.15	0.15	0.34	0.84	1.01	818.61	305.30
4	sd_E__I	0.94	0.56	1.11	0.53	0.06	3.39	1.01	738.04	291.92
5	r_E[1,I]	0.32	0.21	0.77	0.39	-0.64	1.50	1.01	482.44	232.87
6	r_E[2,I]	-0.06	-0.04	0.74	0.35	-1.14	0.98	1.01	520.36	237.73
7	r_E[3,I]	-0.20	-0.16	0.76	0.36	-1.39	0.80	1.01	544.38	239.68
8	lprior	-20.50	-20.50	0.02	0.01	-20.55	-20.49	1.01	736.76	292.70
9	lp__	-158.56	-158.26	2.45	2.36	-163.03	-155.07	1.00	1481.31	1861.92

Table 3: Hierarchical model.

For the pooled model, The Rhat values for all parameters are at 1.00, which is ideal. An Rhat value of 1.00 indicates that between-chain variance is comparable to within-chain variance, suggesting that the chains have mixed well and convergence has been reached.

The ESS values for the main parameters (b_Intercept, b_PAY.0, b_PAY.2) are all above 2000, which is typically considered sufficient for reliable estimates. A higher ESS value indicates that the effective information drawn from the posterior distribution is substantial, reducing the uncertainty of the estimates.

The tail ESS values are also high, particularly for the intercept and PAY_0, indicating that the estimation of the tail areas of the posterior distribution is reliable, which is important for accurate quantile estimation.

The hierarchical model also shows good convergence with Rhat value at 1.01 for all parameters. And the ESS values are between ∼500 to ∼1500 that leads to less precise estimates of the posterior means.

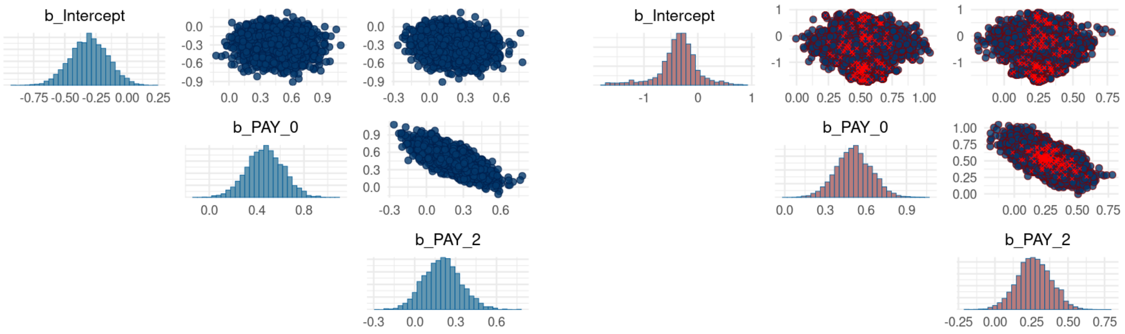
5.2 HMC convergence diagnostics

It can be seen from Figure 2 that the pooled model converges without any issues. The off-diagonal scatter plots show no red points, indicating there were no divergent transitions during sampling, which is a positive sign for convergence. The diagonal histograms show the marginal posterior distributions for each parameter, which appear to be roughly normally distributed, suggesting no issues with skewness or multimodality.

However, the hierarchical model reveals a significant number of red points, indicating a number of divergent transitions. This is a potential sign of issues with convergence, indicating that the sampler may have struggled to explore the posterior distribution effectively.

The red points are scattered across the parameter space, which could suggest that the step size might be too large or that the model may have areas of problematic geometry that are causing the HMC sampler to diverge.

The elongation in scatter plots, especially for b_PAY_2, is more pronounced here, suggesting stronger correlations between some parameters. The presence of divergences in these plots indicates that these correlations may be contributing to the difficulty the sampler is experiencing.



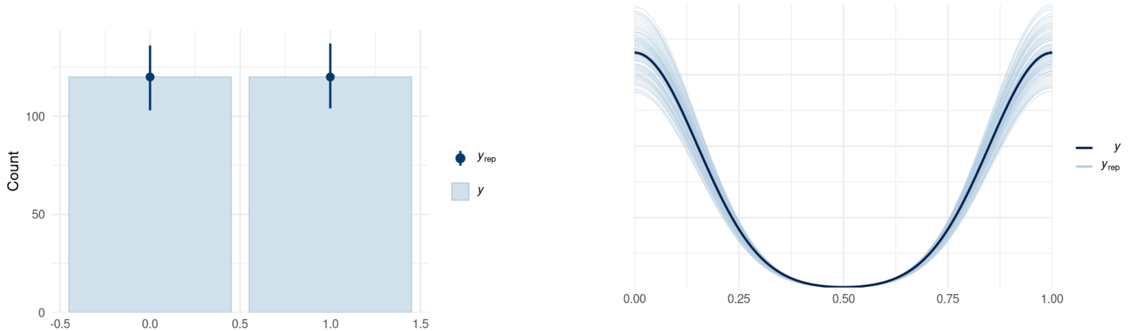
(a) MCMC diagnostics for the pooled model

(b) MCMC diagnostics for the hierarchical model

Figure 2: Comparative MCMC diagnostics for pooled and hierarchical models

6 Posterior predictive checks

6.1 PPC for the pooled model



(a) Posterior Predictive Check for Proportions

(b) Posterior Predictive Check for Predicted Probabilities

Figure 3: Posterior predictive checks using pooled model. Figure (a) presents the comparison of observed counts to the posterior predictions. Figure (b) overlays the observed proportion densities on the predicted proportion densities.

The Figure 3a shows the observed counts of the binary outcomes compared to the counts from replicated datasets drawn from the posterior predictive distribution. The blue bars represent the observed data (y), and the black dots with error bars represent the median and variability (usually 50% and 95% intervals) of the predicted counts (y_rep) from the posterior predictive distribution.

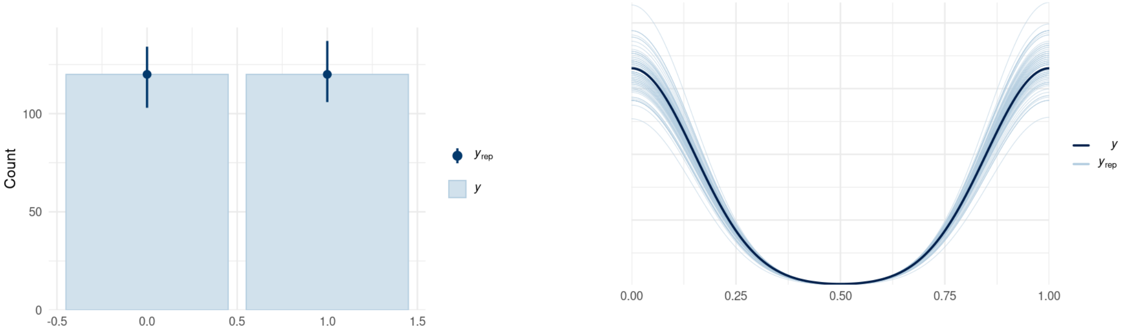
The fact that the observed counts fall within the range of the posterior predictive intervals for both categories suggests that the model has a reasonable fit. It does not appear to be systematically overestimating or underestimating the frequency of either outcome.

Figure 3b overlays the density of the observed data on the densities of several replicated datasets (predictions) from the posterior predictive distribution.

The dark line represents the observed proportions (y), while the lighter lines represent the densities of predicted proportions (y_rep) across different samples from the posterior distribution.

The close alignment of the predicted densities around the observed proportions suggests that the model is capturing the underlying distribution of the data well.

6.2 PPC for the hierarchical model



(a) Posterior Predictive Check for Proportions

(b) Posterior Predictive Check for Predicted Probabilities

Figure 4: Posterior predictive checks using hierarchical model. Figure (a) presents the comparison of observed counts to the posterior predictions. Figure (b) overlays the observed proportion densities on the predicted proportion densities.

We can see in Figure 4 that the hierarchical model behaves similarly as the pooled model. The predicted density follows very closely with the observed data as well as the observed counts fall within the range of the posterior predictive intervals for both categories. Therefore, we can conclude that both model behave similarly and both predict quite well the observe data.

7 Predictive performance assessment

Based on the Table 4, which compares the predictive performance of pooled and hierarchical models using both short data (240 entries) and original data (30,000 entries), several points can be drawn about the practical usefulness of the accuracy reported:

Accuracy Levels: Both models achieve the same accuracy with the original data set (0.80), which is quite high and indicates that either model could be practically useful for making predictions on this data. For the short data, the hierarchical model (0.75) performs slightly better than the pooled model (0.73).

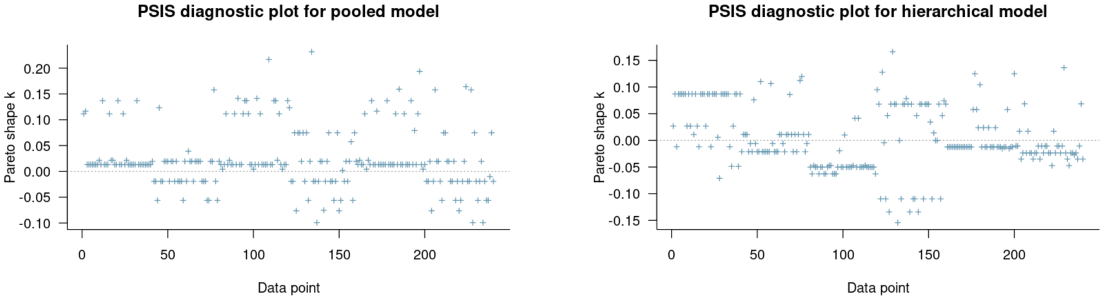
Confidence Intervals (CI): The confidence intervals for the original data are narrower than those for the short data, which is expected due to the larger sample size providing more information and thus greater precision in the estimate of accuracy. The overlap in the 95% CIs for both models with the original data suggests no significant difference in performance between the pooled and hierarchical models at the 0.80 accuracy level.

Practical Usefulness: The noninformative rate (0.7788) for the original data suggests that if a model were to predict 'no default' for all cases, it would be correct approximately 77.88% of the time due to class imbalance. Any predictive model used in this context must significantly outperform this baseline to be considered practically useful.

The reported accuracies of 0.80 for both the pooled and hierarchical models with the original data represent only a slight improvement over the noninformative rate. This raises the question of whether the model is truly capturing the underlying risk factors for default or primarily reflecting the imbalance in the dataset.

Table 4: Predictive Performance of Models with prior $\mathcal{N}(0, 100)$. The last column is the noninformative rate, the ratio of zeros to number of data points.

Model	Data Type	Accuracy	95% CI	Rate
Pooled	Short Data (240 entries)	0.73	(0.668, 0.784)	0.5
Pooled	Original Data (30000 entries)	0.80	(0.806, 0.814)	0.7788
Hierarchical	Short Data (240 entries)	0.75	(0.695, 0.807)	0.5
Hierarchical	Original Data (30000 entries)	0.80	(0.795, 0.804)	0.7788



(a) PSIS diagnostic plot for the pooled model

(b) PSIS diagnostic plot for the hierarchical model

Figure 5: Pareto Smoothed Importance Sampling (PSIS) diagnostics for pooled and hierarchical models

The PSIS diagnostic plots in Figure 5 show the Pareto k-values for each observation in the model, which assess the reliability of the variational inference used in approximating the posterior distribution.

All the Pareto k-values seem to be well below 0.7, which is generally considered good. Values below 0.5 are typically seen as indicating no issues, and all the points in both the pooled model and the hierarchical plot fall under this threshold. Both models do not exhibit signs of problematic importance weights, suggesting that the variational approximations of the posterior distributions are reliable for both models.

8 Sensitivity analysis

It can be observed from Table 4 and Table 5, the model's accuracy on the short data (240 entries) and the original data (30000 entries) are reported alongside their 95% confidence intervals (CI) and

Table 5: Predictive Performance of Models with prior $\mathcal{N}(\log(\frac{0.7}{0.3}), 1)$

Model	**Data Type**	**Accuracy**	**95% CI**	**Rate**
Pooled	Short Data (240 entries)	0.73	(0.673, 0.788)	0.5
Pooled	Original Data (30000 entries)	0.81	(0.804, 0.813)	0.7788
Hierarchical	Short Data (240 entries)	0.77	(0.712, 0.822)	0.5
Hierarchical	Original Data (30000 entries)	0.80	(0.804, 0.813)	0.7788

a non-informative rate. The non-informative rate is calculated as the ratio of zeros to the number of data points, providing insight into the sparsity of the dataset.

From the results, we observe only minor differences in accuracy for the short data set across the two priors, with the normal prior resulting in slightly higher accuracy. However, for the original data set, the accuracy remains the same regardless of the prior used. This might be due to the large size of the dataset, therefore, the choice of prior has a negligible effect on the model's performance. The data's strength and consistency overpower the prior's influence, indicating that the data is the main driver of the results. Moreover, the consistency in accuracy for the original data set across different priors may imply that the model is relatively insensitive to the prior specifications.

Overall, these results indicates that for the given logistic regression models, the prior has a minor impact on the performance with a compact data set and virtually no impact with a full data set.

9 Model comparison

The Table 6 shows that the hierarchical model (fit2) has an elpd_diff of 0, which serves as the baseline, and the pooled model (fit1) has a lower elpd_diff, indicating that it has a worse predictive performance than the hierarchical model by 7.72 points.

The se_diff is the standard error of the difference in ELPD. The pooled model has a se_diff of 12.17, which indicates the uncertainty associated with the elpd_diff.

The LOO information criterion, a measure for model comparison with a penalty for model complexity. The hierarchical model has a lower looic (272.32) compared to the pooled model (287.76), indicating better predictive performance when the complexity is accounted for.

Therefore, the hierarchical model (fit2) shows better predictive performance compared to the pooled model (fit1), as indicated by the higher elpd_loo and the lower looic.

Table 6: Comparison between pooled and hierarchical model

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
fit2	0.00	0.00	-136.16	7.88	4.92	0.43	272.32	15.77
fit1	-7.72	12.17	-143.88	9.39	4.27	0.98	287.76	18.77

10 Discussion of issues and potential improvements

Model Accuracy vs. No Information Rate: Both models' accuracies do not significantly outperform the no information rate. This suggests that the models might be doing little more than predicting the most common class. For credit default prediction, this is particularly concerning because it may lead to a high rate of false negatives, where defaults are not predicted, potentially resulting in substantial financial risk. The similar accuracy levels of both models to the no information rate could also indicate that key predictive variables or interactions are missing from the models, or that the data contains a lot of noise, which the current model structures are not capturing.

Divergences in the Hierarchical Model: The presence of divergent transitions in the hierarchical model is a significant issue as it suggests problems with exploring the parameter space. This can lead to parts of the space being underexplored, potentially resulting in biased estimates or overconfidence in the model's predictions. Divergences might also imply that the model's current parametrization is challenging for the HMC algorithm, possibly due to complex posterior geometry or highly correlated parameters.

Comparative Performance of Models: Despite the hierarchical model's ability to account for group-level variability, it does not significantly outperform the pooled model in terms of accuracy. This could mean that the added complexity of the hierarchical model is not translating into better predictive performance for the data at hand, or that both models are equally limited by other factors such as data quality or feature selection.

To address divergences in the hierarchical model, consider reparameterizing the model. Non-centered parameterizations are often helpful for hierarchical models as they can reduce parameter correlations and improve sampler efficiency.

Because the dataset is imbalanced, techniques such as resampling, synthetic data generation (e.g., SMOTE) can also be used to further improve the model predictability as well.

11 Conclusion

In summary, our findings indicate that both models perform comparably in terms of accuracy. However, the practical utility of these models is called into question by their performance relative to the no-information rate, suggesting that the predictive power of both models is not substantially better than a naive classification by the most frequent outcome.

The hierarchical model, despite its sophisticated approach to incorporating educational background, must be approached with caution due to the presence of divergent transitions during the sampling process. These divergences highlight potential issues in the model estimation process, which could impact the reliability of its predictions.

Furthermore, the impact of priors on the model was found to be minimal. Given the large size of the dataset, the data's inherent patterns appear to overshadow the influence of the priors, rendering the choice between non-informative and weakly informative priors less consequential in the final model outcomes.

12 Self-reflection

During the project, the ability to perform the entire Bayesian workflow was acquired. These involve building models, selecting certain values for posterior predictive checks, and making inferences from simulations performed, in addition to utilising a variety of libraries relevant to Bayesian analysis.

Reflecting on this project, we've learned the importance of aligning model complexity with actual performance gains. While the hierarchical model promised a more tailored approach by considering educational background, it also introduced complexities that challenged its stability, as evidenced by the divergent transitions, underscoring the need for careful model assessment. The overwhelming influence of a large dataset on the priors highlighted the power of data in driving model results. This experience has taught us the value of rigorous model evaluation against simple benchmarks and has sharpened our approach to interpreting model diagnostics and assessing their real-world utility.

References

[1] https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

[2] Islam, S.R., Eberle, W., Ghafoor, S.K. (2018). Credit Default Mining Using Combined Machine Learning and Heuristic Approach. ArXiv, abs/1807.01176.

[3] Fung, C., Koerner, J., Grant, S., Beschastnikh, I. (2018). Dancing in the Dark: Private Multi-Party Machine Learning in an Untrusted Setting. ArXiv, abs/1811.09712.