ST308 Project Report

Candidate Number: 22370

Date Submitted: 06/05/2021

1 Empirical problem description & EDA

1.1 Dataset description

The dataset contains 5000 rows of cross-sectional data on different characteristics of customers from a bank loan promotion campaign. Our aim is to analyse the data and propose a Bayesian classification model that will help a US bank to predict the likelihood of a customer buying a personal loan.

The dataset contains the following variables, whose description can be found in sheet 1 of the dataset's Excel document:

Independent Categorical variables: ZIP Code, Family, Education, Securities Account, CD Account, Online, Credit Card.

Independent Continuous variables: Age, Experience, Income, CCAvg, Mortgage

Dependent variable: Personal Loan; '1' for customers who accepted the personal loan offered in the previous campaign, and '0' otherwise.

1.2 Suitability of the dataset for a Hierarchical model

The dataset contains a mixture of interesting numerical and categorical features which makes the analysis very interesting. One of the main tasks in this coursework, is to fit a Hierarchical model to address the problem using the available data. What makes this dataset suitable for this task, is the ZIP Code feature of each individual customer. However, the cardinality of the ZIP Code feature is extremely large, and this would make the analysis very hard. Therefore, in order to deal with this issue while also maintaining the structure of the dataset, we have used the 'uszipcode' library to replace the ZIP Code variable with a County variable which has a significantly lower cardinality. The result can be visualized in the barplot below, which shows the number of customers in each County, and how many of them have accepted the personal loan in the previous campaign.

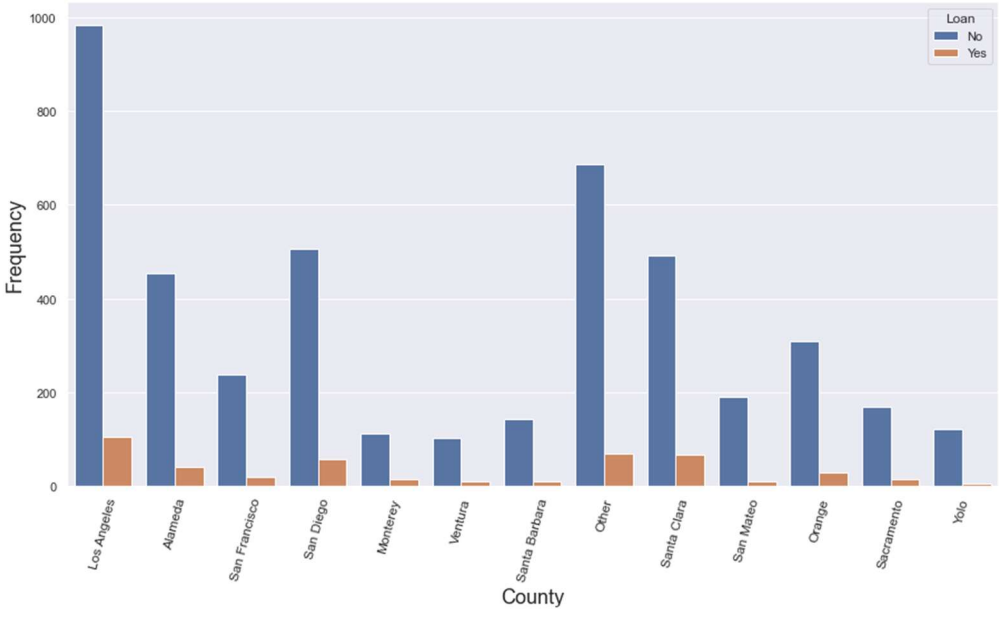

Figure 1: Barplot of the County feature.

1.3 Exploratory Data Analysis (EDA)

We start this section by briefly describing the data cleaning process which is necessary to bring the dataset to a form which can be used in the modelling part. We refer to figure 1 of the appendix and see that there are no missing values in our dataset. However, we notice that some rows have a value for the Experience feature which is negative. By carefully observing the Age of those individuals, we can deduce that this issue can be easily fixed by taking the absolute value of the Experience feature.

The aim of the EDA part is to develop a better understanding of the unique characteristics of the dataset, and develop some intuition as to which features might be helpful for our model. For example, by observing the figure below, the histogram and boxplot of the Income feature clearly suggest that the distribution of income is positively skewed with extreme outliers on the right tail of the distribution. Moreover, by looking at the income distribution between customers who previously accepted the loan and those who did not, we might suspect that this feature will be of particular importance for our models.

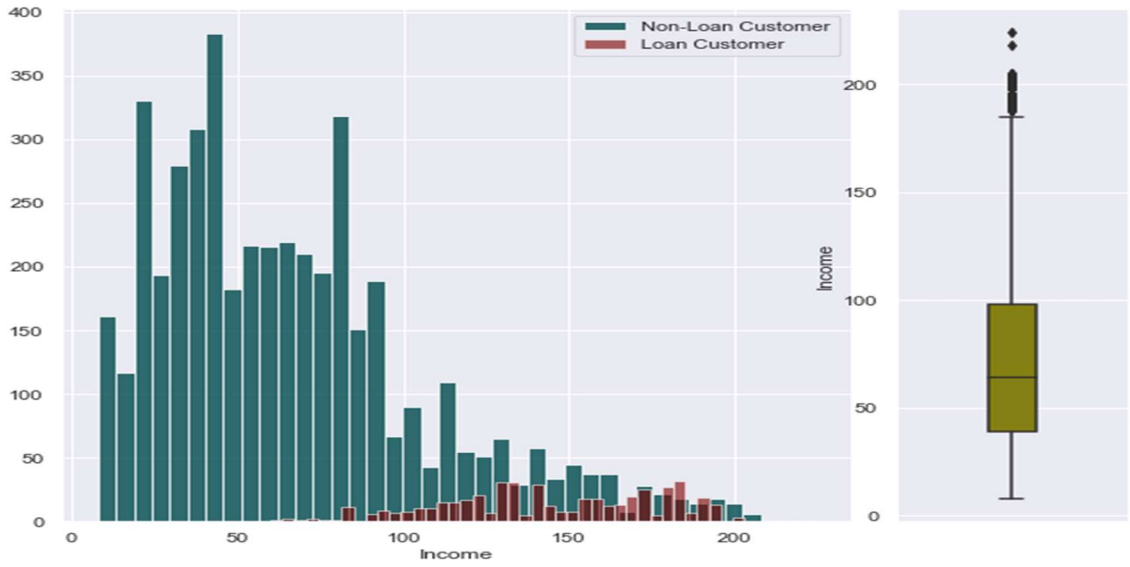

Figure 2: Histogram and boxplot of the Income feature.

By analysing other numerical features in our dataset in a similar manner, we can see that, amongst other observations, the majority of customers are aged between 30 and 60, and most of them do not have a house mortgage. Moving to categorical features, we notice something interesting, more customers with larger family sizes seem to have accepted the personal loan in the previous campaign. In addition, almost all customers who had a certificate of deposit (CD) account, had accepted the personal loan. In order to improve the fit of the classification models, we highlight the importance of dealing with the extreme outliers in the Mortgage and CCAvg features. Dropping observations with z-scores of more than 4 results to a loss of less than 1% of the data, and so we can proceed with dropping them. In figure 2 of the appendix, we note the

essentially perfect correlation between the Age and Experience features, and drop the latter to avoid introducing multicollinearity between our predictors. For completeness, we also mention that even though it would be a good idea to transform right-skewed continuous predictors (such as Income and CCAvg), we omit that for ease of interpretation of the model coefficients.

2 Competing classification models

2.1 Bayesian logistic regression

In this section, we address the first part of the project which is to come up with a Bayesian logistic regression model that will allow us to better understand which variables seem to be more important for predicting whether a customer is likely to accept the personal loan. Let us mention that this dataset is well suited to a Bayesian data analysis because being able to quantify uncertainty when analysing the likelihood of whether a customer will accept the loan or not could be very helpful for the bank's marketing department when devising their campaign strategy.

An advantage of the Bayesian logistic regression model in contrast to the frequentist logistic regression model fitted via Maximum Likelihood Estimation (MLE), is that it allows us to incorporate any prior information we might have about the model parameters of interest based on domain knowledge and common sense. Below, we specify the model for our data, and describe the different models used throughout our analysis.

Let Y_i be a Bernoulli random variable that takes the value 1 if the i^{th} customer accepts the personal loan, and 0 otherwise. Then, $Y_i \sim Bernoulli(\pi_i)$ where:

$$\text{logit}(\pi_i) = a + \sum_{j=1}^{p} \beta_j X_j^i$$

where p is the number of predictor variables in our dataset, X_j^i is the value of the j^{th} predictor of the i^{th} customer, and the a and β_js are the intercept and slope coefficients to be estimated.

Finally, the probability that the i^{th} customer accepts the personal loan is given by the following expression:

$$\pi_i = \frac{1}{1 + \exp\left(-a - \sum_{j=1}^{p} \beta_j X_j^i\right)}$$

As baseline models, we firstly try fitting a logistic regression using MLE, and a Bayesian logistic regression model using the Laplace approximation for the posterior distribution of the parameters. To do this, we choose a unit information prior for our parameters vector. That is, we assign the prior $\boldsymbol{\beta} \sim N(\mathbf{0_{p+1}}, \frac{1}{n} X^T X)$, where $\mathbf{0_{p+1}}$ is a (p+1)-dimensional zero vector and X represents the data matrix of our training set. The intuition behind using a unit information prior is that we try to 'let the data speak for itself', and so we aim for a prior on our parameters that will add as much information as a single observation. In theory, using a Normal approximation for our parameters' posterior distribution should be relatively accurate, given that the size of our training set is sufficiently large. In fact, in figure 3 of the appendix, we see that in terms of predictions on the test set, both the MLE and the Bayesian (using the Laplace approximation) logistic regression models perform almost equally well. As we should expect with a unit information prior, our parameter estimates in both models are relatively close to each other, and this explains why we observe very similar balanced accuracy and ROC-AUC scores.

Moving on to the main model of this section, and keeping the same notation we introduced earlier, we now assign different priors on our parameters. For the intercept term a, we assign a $Cauchy(0\,, 2)$ prior distribution. The reason behind doing this is that we want to use the heavy tails property of the Cauchy distribution to reflect the uncertainty that we have about the value of the a parameter. Hence, a Cauchy prior is preferable over a Normal prior for example. Moving on to the p-dimensional slope coefficients vector $\boldsymbol{\beta}$, we choose a multivariate prior that should reflect the fact that almost all of the predictors are essentially uncorrelated with each other, as we observed in the correlation matrix previously. To achieve this, we need to have a diagonal covariance matrix. Hence, we choose to go for a weakly informative multivariate Normal prior, $N(\mathbf{0_p}\,,\ 100\, I_p)$, where $\mathbf{0_p}$ is a p-dimensional zero vector and I_p is the pxp identity matrix.

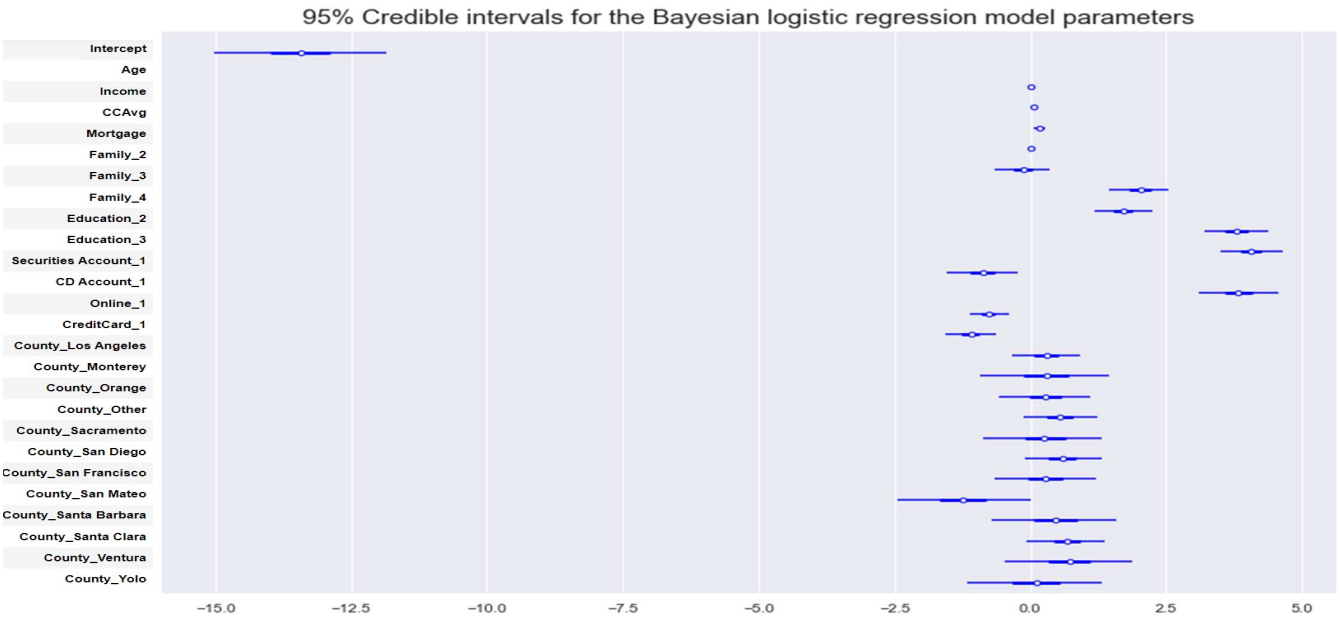

Figure 3: 95% credible intervals for the Bayesian logistic regression model.

By observing figure 3, through the 95% credible intervals, we get a probabilistic sense of our posterior parameter estimates, as opposed to the frequentist confidence intervals which cannot be interpreted in the same way. For example, we can confidently infer from the relatively narrow credible intervals that customers with more advanced education levels, and customers with larger families have a higher chance of accepting the loan offer. In contrast, the extremely narrow credible intervals for the Age and Mortgage predictors, with posterior mean estimates which are essentially zero, really pose the question of whether these variables are of any practical usefulness in our model. Something important to point out is that most of the credible intervals for the County variables contain the value zero. This might be because the fixed effects model does not have high enough power to detect significant changes for each County level. Lastly, we note that the intercept term acts as a collector and represents the reference level of each categorical variable.

We now refer to figure 4 of the appendix, where we provide a summary of the fitted model coefficients. Comparing them with the coefficients of the logistic regression model via MLE, we notice that the coefficients of most predictors are roughly similar. However, something that stands out is that some of the coefficients for the County dummy variable predictors have significant differences compared to those of the MLE logistic regression model. This strongly points us towards considering a Hierarchical logistic regression model which may prove to better fit our data.

2.2 Hierarchical logistic regression models

In this part we present the two Hierarchical logistic regression models considered in the analysis and provide justification as to which of the two is the most appropriate model for our data. We try to incorporate information from the County of each customer, and hence attempt to vary the model coefficients according to the County levels.

2.2.1 Random intercept model

We start by describing our random intercept model for the loan acceptance probability of the i^{th} customer in the j^{th} County level as follows:

$$\text{logit}(\pi_{ij}) = a_j + \sum_{k=1}^{p} \beta_k X_k^{ij} \qquad with\ j = 1, 2, \dots, 13\ and\ i = 1, 2, \dots, n_j$$

where n_j is the number of customers in the j^{th} county, a_j is the intercept term in the j^{th} County level, and X_k^{ij} is the value of the k^{th} predictor of the i^{th} customer in the j^{th} County level.

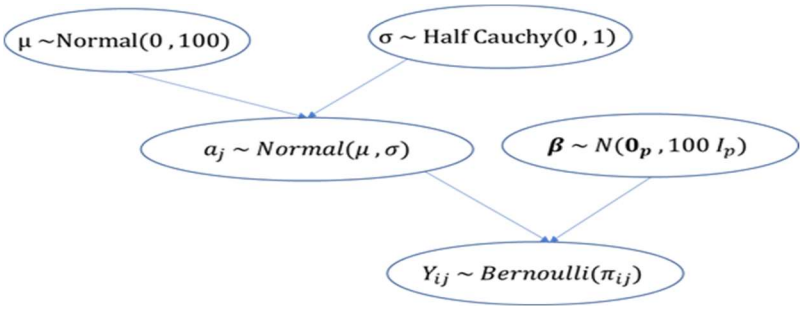

Figure 4: Directed acyclic graph (DAG) summary of the random intercept model.

2.2.2 Random intercept and slope model

By considering the results of the Bayesian logistic regression model it seems reasonable to also try to add random effects in the slope coefficients of predictors that seem to be important in our model. As the number of Counties is large, we only consider a model with a random slope for a single predictor to prevent the number of parameters from growing very large. Based on the previous credible interval results, choosing the predictor CCAvg seems a good idea. Hence, similar to our above model description, we have:

$$\text{logit}(\pi_{ij}) = a_j + \gamma_j Y^{ij} + \sum_{k=1}^{p-1} \beta_k X_k^{ij} \qquad with\ j = 1, 2, \dots, 13\ \ and\ i = 1, 2, \dots, n_j$$

where γ_j is the CCAvg slope coefficient in the j^{th} County level, Y^{ij} is the CCAvg value of the i^{th} customer in the j^{th} County level, and the β coefficients are for all predictors except CCAvg.

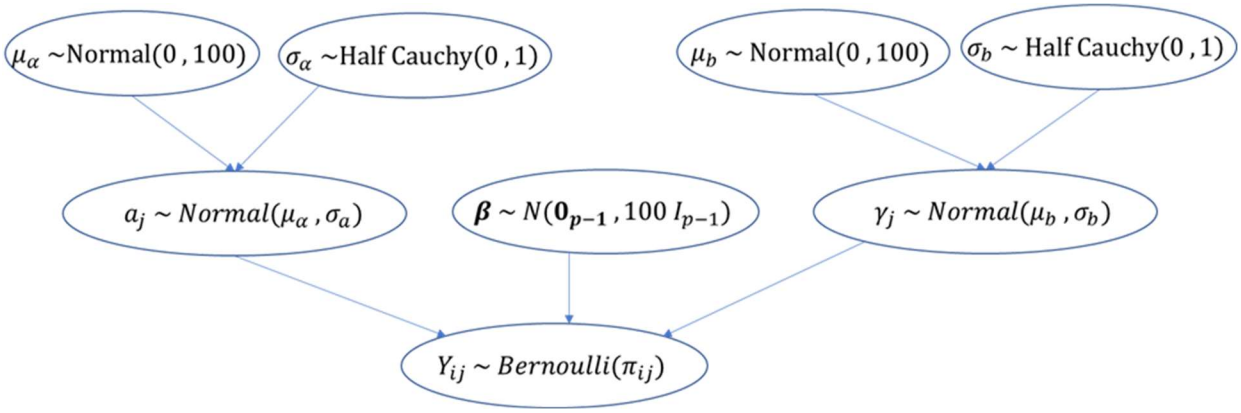

Figure 5: DAG summary of the random intercept and slope model.

2.3 MCMC Diagnostics & Model selection

Before analysing our competing models, we firstly need to ensure that the quality of the samples generated using the Hamiltonian MCMC algorithm is sufficient to provide accurate approximations of the posterior distributions. From the Stan summary outputs in the Jupyter notebook, we see that for all three Bayesian models, all but one of the 'Rhat' values are not larger than 1.02, and the 'n_eff' values are large enough. Furthermore, let us also examine graphically the effectiveness of the chain mixing using the graphs below.

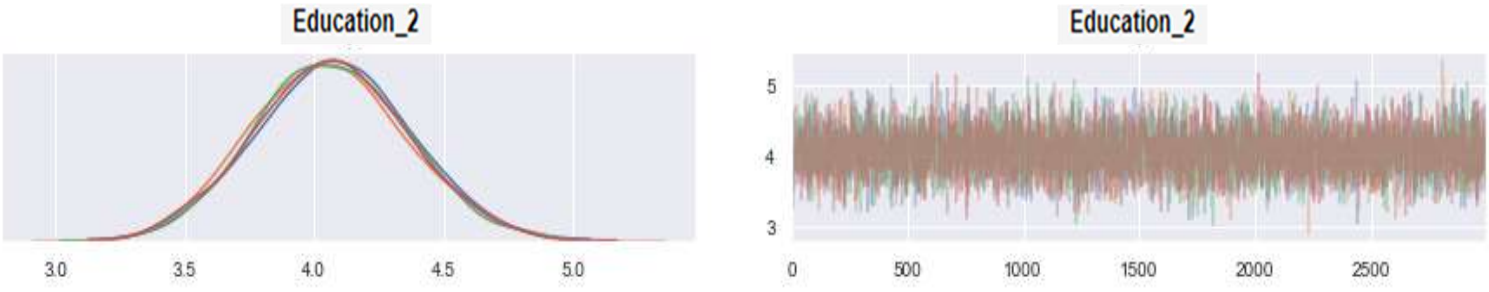

Figure 6: Slope coefficient posterior distribution and MCMC trace plot of the Bayesian logistic regression model.

By noting the thick-haired chains in the right plot, we are confident that the chains have converged and the MCMC algorithm is efficiently exploring the parameter space without showing any odd patterns. Also, there does not appear to be any issue with the chains being highly correlated. With regards to the Hierarchical models, MCMC diagnostics seem to be satisfactory as well. Concluding, after making sure that our model parameter posterior estimates can be trusted, we can proceed to comparing the models and interpreting them.

For the remainder of this section, we briefly talk about the criteria according to which we select the model that seems to best fit our data, having in mind Bayesian Occam's razor. We aim for a model which is complex enough to adequately fit our data while also keeping the model as parsimonious as possible. Since Bayesian models are well-known for naturally handling prediction tasks using their predictive distribution, we compare our models according to their balanced accuracy and AUC scores on the test set. As a side note, since our data is quite imbalanced in terms of the target variable, a metric that considers both sensitivity and specificity is more appropriate, which justifies our choice for the balanced accuracy score. We refer to figure 5 of the appendix and note that our random intercept Hierarchical model performs only marginally better compared to the MLE and Bayesian logistic regression model in terms of AUC. However, the results are extremely similar which renders us unable to select the best model based on test set prediction performance.

Continuing from above, we carry out Bayesian model selection using Leave-one-out (LOO) cross-validation for estimating the out-of-sample prediction accuracy of our models. By defining the log-likelihood of each model in our Stan code, we are able to provide a table which summarises the three Bayesian models considered in section 2.

	rank	loo	p_loo	d_loo	weight	se	dse	warning	loo_scale
Random intercept model	0	990.716	17.5883	0	0.656023	62.3171	0	False	deviance
Random intercept & slope model	1	992.519	19.8812	1.80308	0.276366	62.4795	1.10365	False	deviance
Bayesian LogReg	2	998.318	25.5306	7.6023	0.0676108	63.009	4.46412	False	deviance

Figure 7: Bayesian model selection using Leave-one-out cross-validation.

We observe that the Hierarchical models clearly provide a better fit for our data indicated through a lower value in the 'Loo' column. In addition, the difference between these two models seems to be small, but keeping in mind model parsimony, we decide to go with the random intercept Hierarchical model.

3 Model Interpretation and conclusion

We start this section by interpreting the coefficients of our chosen random intercept Hierarchical model. As the number of parameters in this model is quite large, we do not explicitly write down the fitted model, although a detailed description of the model coefficients can be found in figure 6 of the appendix. As a first observation, we note that the slope coefficients are relatively similar to the MLE logistic regression model. This probably arises from the use of weakly informative priors for our parameters. Prior distributions are an important part of any Bayesian analysis, as they can largely influence the final model results and conclusions.

Therefore, our analysis could have benefitted from domain knowledge of the bank's marketing team on their customers' characteristics.

We start by explaining how to interpret the coefficients of continuous predictors, using CCAvg as an example. Taking the CCAvg posterior mean to be 0.17, we see that keeping everything else fixed, for a \$1000 increase in monthly credit card average spending, the odds of a customer accepting the personal loan are multiplied by $e^{0.17} \approx 1.185$. Equivalently, a \$1000 increase in monthly credit card average spending increases the odds of a customer accepting the personal loan by 18.5%. A similar interpretation follows for other continuous predictors as well. Moving on to categorical predictors, we use the Family_2 categorical dummy variable predictor as an example. A posterior mean coefficient estimate of 1.64 implies that keeping everything else fixed, the odds of a customer of family size 2 accepting the personal loan decrease by approximately 12.2%, relative to the Family_1 reference group.

Analysing the posterior distributions of the varying intercept coefficients, we can see that there are some differences in them which suggests that the model indeed detects significant differences for customers in different counties. The alpha[9] and alpha[11] coefficients correspond to the San Mateo and Santa Clara counties respectively, and have posterior mean estimates of -13.13 and -12.76. This means that the odds of a Santa Clara resident accepting the personal loan are 44.8% higher compared to a San Mateo resident with the exact same characteristics. Thus, from the analysis we have conducted, and using a similar approach to analyse the remaining varying intercept coefficients, we can conclude that the bank's marketing team can really improve the effectiveness and efficiency of their campaign by incorporating information about a customer's county of residence. Finally, we mention that based on the credible intervals of the posterior estimates, all of the predictors except Age and Family_2 will be of high importance to the bank.

The relatively large estimate of the 'sigma' parameter in figure 8 below reveals that there is significant between group variability for the county-varying intercepts, justifying the use of a Hierarchical model to address this problem.

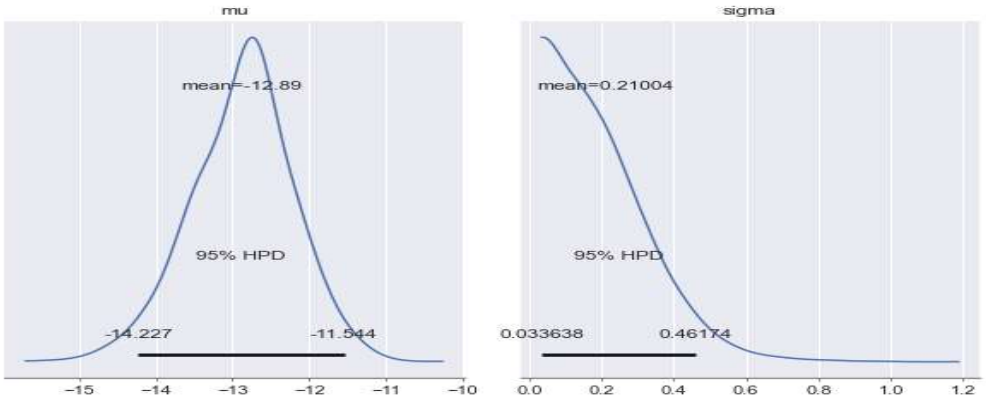

Figure 8: Posterior distribution and 95% credible intervals for the 'mu' and 'sigma' parameters.

As a concluding remark, we state that the results of this analysis show that our Bayesian Hierarchical model does not provide significant evidence in terms of better predictive performance compared to the frequentist logistic regression model. Nevertheless, being able to quantify the uncertainty in the parameter estimates is one of the powerful things about Bayesian modelling.

Appendix

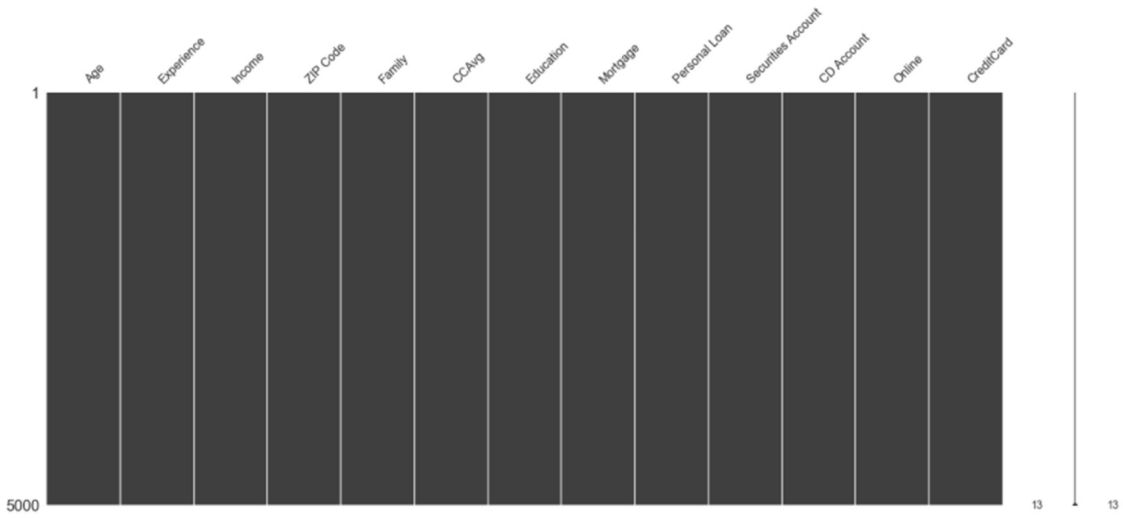

Figure 1: Visualizing missing values in the dataset using the 'msno' library.

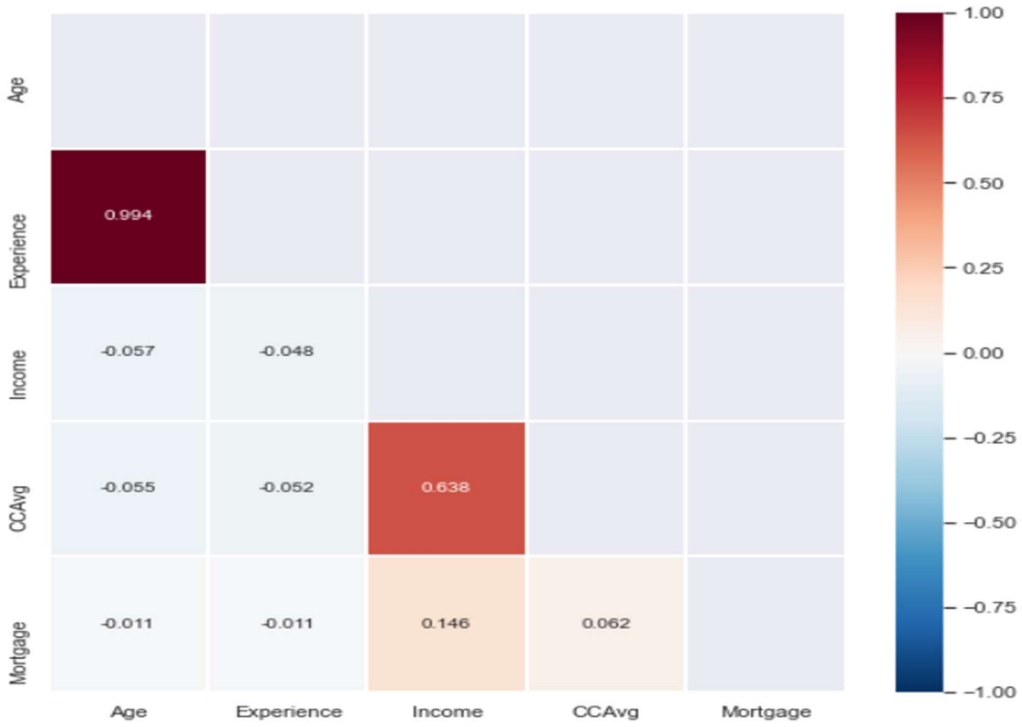

Figure 2: Correlation matrix plot for the numerical features.

```
MLE test set balanced accuracy: 0.8172 and AUC: 0.9640
Bayesian LogReg using Laplace test set balanced accuracy: 0.8095 and AUC: 0.9642
```

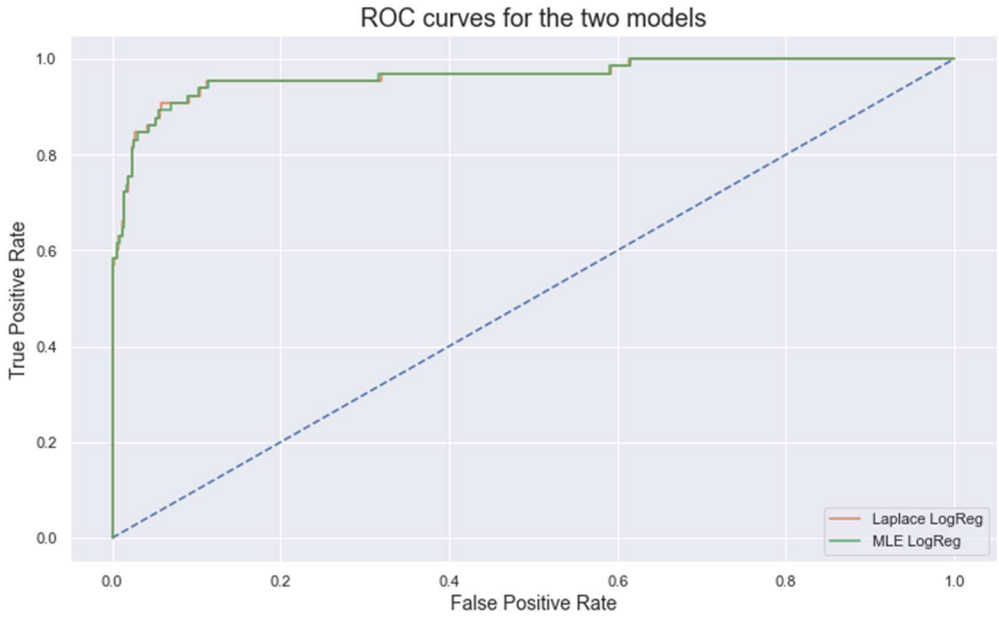

Figure 3: ROC curves for the MLE and Bayesian (using a Laplace approximation) logistic regression models.

	mean	se_mean	sd	2.5%	97.5%	n_eff	Rhat
Intercept	-13.44	0.01	0.8	-15.05	-11.9	3463	1.0
Age	3.5e-3	8.0e-5	7.8e-3	-0.01	0.02	9440	1.0
Income	0.06	4.8e-5	3.5e-3	0.06	0.07	5380	1.0
CCAvg	0.17	5.1e-4	0.05	0.06	0.27	10054	1.0
Mortgage	1.6e-3	5.3e-6	7.3e-4	2.1e-4	3.1e-3	18599	1.0
Family_2	-0.14	2.8e-3	0.26	-0.65	0.36	8344	1.0
Family_3	2.03	3.4e-3	0.28	1.49	2.59	6678	1.0
Family_4	1.71	3.3e-3	0.27	1.19	2.25	6764	1.0
Education_2	3.8	4.0e-3	0.31	3.22	4.42	5803	1.0
Education_3	4.07	4.1e-3	0.3	3.5	4.67	5252	1.0
Securities Account_1	-0.88	3.4e-3	0.33	-1.55	-0.24	9604	1.0
CD Account_1	3.83	4.5e-3	0.38	3.1	4.59	7150	1.0
Online_1	-0.77	1.8e-3	0.19	-1.14	-0.41	10453	1.0
CreditCard_1	-1.1	2.5e-3	0.24	-1.58	-0.64	8890	1.0
County_Los Angeles	0.3	5.0e-3	0.33	-0.32	0.96	4315	1.0
County_Monterey	0.28	7.0e-3	0.61	-0.95	1.44	7740	1.0
County_Orange	0.28	5.8e-3	0.43	-0.57	1.11	5387	1.0
County_Other	0.54	5.0e-3	0.35	-0.14	1.23	4813	1.0
County_Sacramento	0.26	6.7e-3	0.56	-0.86	1.33	6845	1.0
County_San Diego	0.59	5.3e-3	0.36	-0.1	1.31	4669	1.0
County_San Francisco	0.28	6.1e-3	0.48	-0.67	1.2	6106	1.0
County_San Mateo	-1.25	7.3e-3	0.63	-2.52	-0.03	7429	1.0
County_Santa Barbara	0.46	7.0e-3	0.59	-0.71	1.58	6930	1.0
County_Santa Clara	0.67	5.5e-3	0.37	-0.04	1.39	4477	1.0
County_Ventura	0.71	6.9e-3	0.6	-0.54	1.84	7558	1.0
County_Yolo	0.09	7.2e-3	0.64	-1.22	1.27	8070	1.0

Figure 4: Stan summary of the estimated coefficients of the Bayesian logistic regression model.

```
MLE test set balanced accuracy: 0.8172 and AUC: 0.9640
Bayesian Logistic regression test set balanced accuracy: 0.8095 and AUC: 0.9643
Random intercept Hierarchical Logistic regression test set balanced accuracy: 0.8102 and AUC: 0.9649
```

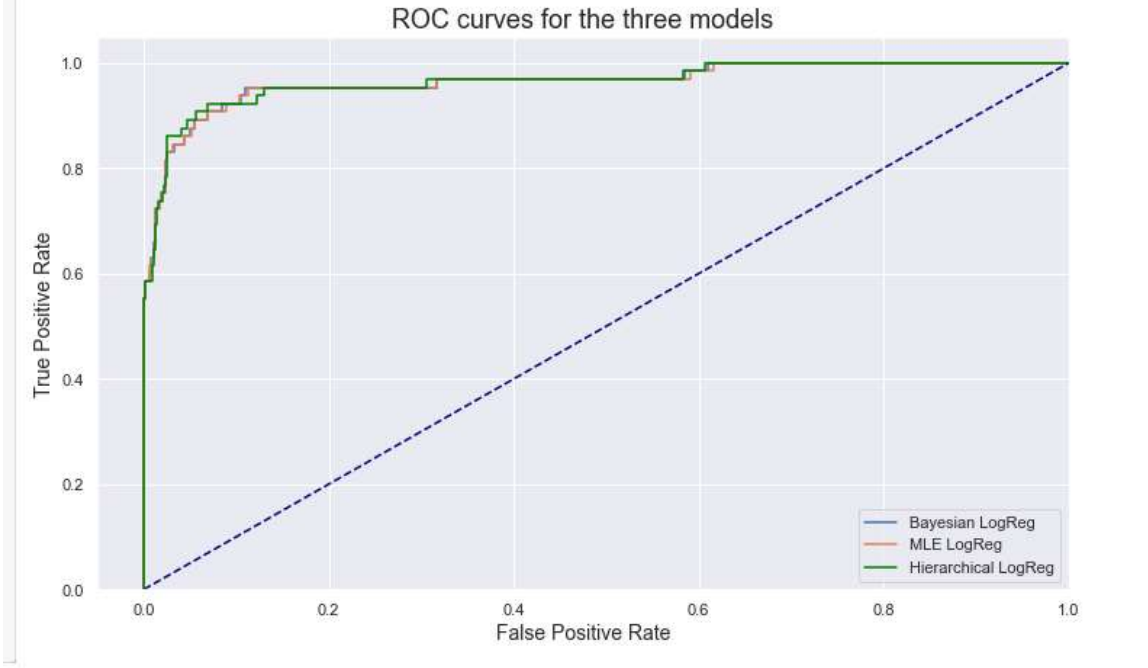

Figure 5: Balanced accuracy and AUC score comparison between the three models.

	mean	se_mean	sd	2.5%	97.5%	n_eff	Rhat
alpha[1]	-13.01	0.02	0.7	-14.52	-11.69	1193	1.01
alpha[2]	-12.9	0.02	0.68	-14.34	-11.62	1224	1.01
alpha[3]	-12.89	0.02	0.7	-14.38	-11.55	1291	1.01
alpha[4]	-12.9	0.02	0.7	-14.38	-11.6	1243	1.01
alpha[5]	-12.8	0.02	0.68	-14.24	-11.51	1246	1.01
alpha[6]	-12.89	0.02	0.7	-14.37	-11.55	1261	1.01
alpha[7]	-12.79	0.02	0.68	-14.21	-11.5	1248	1.01
alpha[8]	-12.9	0.02	0.69	-14.36	-11.57	1267	1.01
alpha[9]	-13.13	0.02	0.74	-14.72	-11.77	1117	1.01
alpha[10]	-12.86	0.02	0.7	-14.33	-11.52	1316	1.01
alpha[11]	-12.76	0.02	0.68	-14.19	-11.46	1265	1.01
alpha[12]	-12.83	0.02	0.7	-14.31	-11.48	1307	1.01
alpha[13]	-12.91	0.02	0.71	-14.43	-11.59	1290	1.01
Age	4.4e-3	2.8e-4	7.7e-3	-0.01	0.02	766	1.01
Income	0.06	7.8e-5	3.2e-3	0.06	0.07	1730	1.0
CCAvg	0.17	1.0e-3	0.05	0.07	0.27	2361	1.0
Mortgage	1.7e-3	1.8e-5	7.1e-4	3.0e-4	3.1e-3	1606	1.0
Family_2	-0.13	0.02	0.27	-0.64	0.39	144	1.03
Family_3	2.02	0.02	0.28	1.48	2.52	214	1.02
Family_4	1.64	0.01	0.26	1.11	2.14	407	1.01
Education_2	3.72	6.8e-3	0.29	3.17	4.31	1775	1.01
Education_3	3.98	7.8e-3	0.28	3.45	4.55	1277	1.01
Securities Account_1	-0.89	0.01	0.33	-1.55	-0.23	1081	1.0
CD Account_1	3.76	6.4e-3	0.36	3.07	4.48	3149	1.0
Online_1	-0.75	4.7e-3	0.18	-1.11	-0.39	1502	1.0
CreditCard_1	-1.03	9.7e-3	0.24	-1.52	-0.6	595	1.01
mu	-12.89	0.02	0.67	-14.31	-11.62	1195	1.01
sigma	0.21	9.1e-3	0.14	0.04	0.53	223	1.02

Figure 6: Random intercept Hierarchical model parameter estimates.