Portuguese bank marketing campaign with GLM

Vittoria Cassola and Andrea Sciortino

2024-07-14

Introduction

This statistical analysis investigates data from a Portuguese banking institution to evaluate the effectiveness of marketing campaigns aimed at promoting subscriptions to bank term deposits. The dataset includes a binary response variable indicating whether a client subscribed to a term deposit, and a variety of covariates encompassing demographic information (such as age and marital status), financial details (including account balance and existing loans), and specifics of the marketing campaign (like the number of contacts made, duration of calls, and prior contact history). In preparing the data for analysis, one significant decision was to omit some variables due to their extensive range of categories, which would have added unnecessary complexity to the study.
We aim to implement a Bayesian approach to fit a model that highlights the key factors influencing clients' decisions to subscribe. Bayesian logistic regression is a tool that allows the combination of prior believes on the parameters with the information coming from the data. The goal is to provide a robust probabilistic framework that identifies significant predictors, offering actionable insights to optimize future marketing strategies and improve campaign effectiveness. The model can be implemented for predictive purposes, to assess whether the campaign will be successful on a client with known characteristics.

GLM with Bayesian approach

The dataset comprises 4521 observations, with 521 resulting in a positive response. We split the data into two subsets: 90% for training and 10% for testing. The training set will be utilized to fit the models, while the test set will be used to evaluate their performance. This approach ensures that our model is validated on unseen data, providing an accurate measure of its predictive capabilities.

Variable selection

We perform variable selection in the context of a generalized linear model (GLM) with a binary response variable to identify which predictors significantly influence whether clients subscribe to a bank term deposit. This process is essential for constructing a parsimonious model that accurately predicts outcomes while avoiding overfitting. By selecting the most relevant predictors, we aim to improve model interpretability and predictive performance.

We introduce a binary vector of dimension (p,1) $\boldsymbol{\gamma}$ such that
$\gamma_j = 1$ if X_j is included in the model
$\gamma_j = 0$ if X_j is not included in the model
The model, with the addition of $\boldsymbol{\gamma}$ treated as a parameter, becomes

$$E(Y|\boldsymbol{x}) = h(\gamma_1\beta_1 X_1 + \ldots + \gamma_p\beta_p X_p)$$

With priors:

$$\gamma_j \sim Bern(w)$$

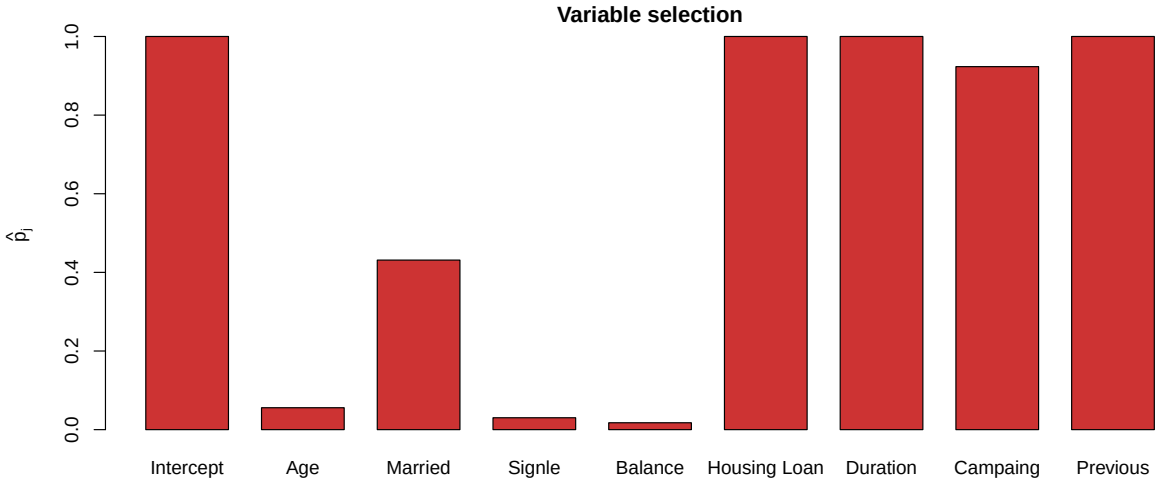
$$\beta_j|\gamma_j = 1 \sim N(\beta_0 j, \sigma_0 j^2)$$

and a resulting joint prior called spike and slab prior

$$p(\beta_j, \gamma_j) = (1-w)\delta_0 + wdN(\beta_j|\beta_0 j, \sigma_0 j^2)$$

We assign to w a non informative prior: $w \sim Beta(1,1)$
From the obtained marginal posterior distribution on γ_j (which is a sequence of 0-1 values) we estimate the posterior probability of inclusion of predictor X_j as

$$\hat{p}X_j = 1/S \sum_{s=1}^{S} \gamma_j^s$$



The plot illustrates striking differences in the posterior probabilities of variable inclusion across our model. Variables such as housing loan, duration of phone calls, campaign contacts, and previous contacts show robust support with probabilities exceeding 0.9, suggesting strong contributions to predicting client subscription behavior. Conversely, age, balance, and marital status categories exhibit posterior probabilities below 0.5, indicating less substantial impacts on the model outcomes. We will exclude variables with low posterior probabilities of inclusion. The remaining variables will be retained for inclusion in the model. This selective approach ensures that our model focuses on the most influential predictors.

Bayesian Logistic Regression

Bayesian logistic regression extends traditional logistic regression by incorporating prior believes about the parameters and updating them with information coming from data to form posterior distributions. Here is a detailed procedure:

- assume the response variable $Y \in (0,1)$ such that $Y_i|\pi_i \sim Ber(\pi_i)$ iid with $i = 1, ..., n$ and

$$\pi_i = h(\boldsymbol{\beta}^T \boldsymbol{x_i})$$

- chose as the inverse-link function $h()$ for $\eta = \boldsymbol{\beta}^T \boldsymbol{x}_i$ the inverse logit

$$h(\eta) = \frac{e^\eta}{1 + e^\eta}$$

We define the logistic regression model with prior on the parameters, $\beta_j \sim \mathcal{N}(\beta_{0j}, \sigma^2_{0j})$ independent.

Once we obtained the posterior sample of the parameters we can investigate the distributions or posterior predictive distribution of subscribing the deposit for a new observation.

Metropolis-Hastings

Due to the nature of the response variable, the prior is not conjugate to the model so, to approximate the posterior distribution, we need to implement a Metropolis-Hastings algorithm:

- Start from initial value $\beta^{(s)} = (\beta_1^{(s)}, \ldots, \beta_p^{(s)})$.

Then for $s = 1, \ldots, S$ and for $j = 1, \ldots, p$:

- Propose: $\beta_j^* \sim q(\beta_j \mid \beta_{-j}^{(s)})$.
- Compute:

$$r_j = \textit{Posterior Ratio } (\beta_j^*, \beta_j^{(s)} | y) \ \cdot \ \textit{Proposal Ratio } (\beta_j^{(s)}, \beta_j^*)$$

- Update:

$$\beta_j^{(s+1)} = \begin{cases} \beta_j^*, & \text{with probability } \min(1, r_j) \\ \beta_j^{(s)}, & \text{with probability} 1 - \min(1, r_j) \end{cases}$$

where $\beta_j^{(s)} = [\beta_k^{(s)}, k \neq j]$

At the end, we obtain the sequence $\boldsymbol{\beta}_j = \{\beta_j^{(1)}, \ldots, \beta_j^{(S)}\}$ for $j = 1, ..., p$. In this case we use adaptive proposal where both mean and variance change across iteration:

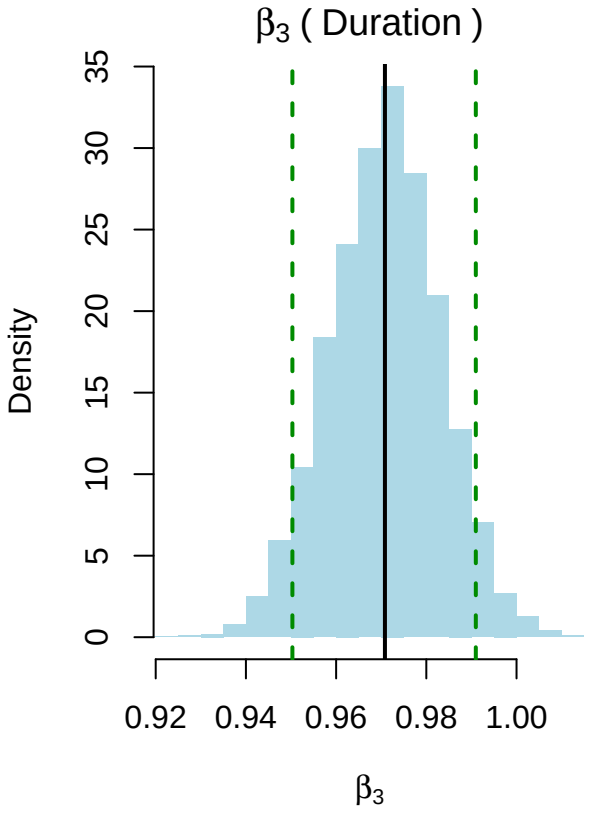
$$q(\beta_j^* | \beta_j^{(s)}) = N(\beta_j^* | m_j^{(s)}, s_j^{2(s)})$$

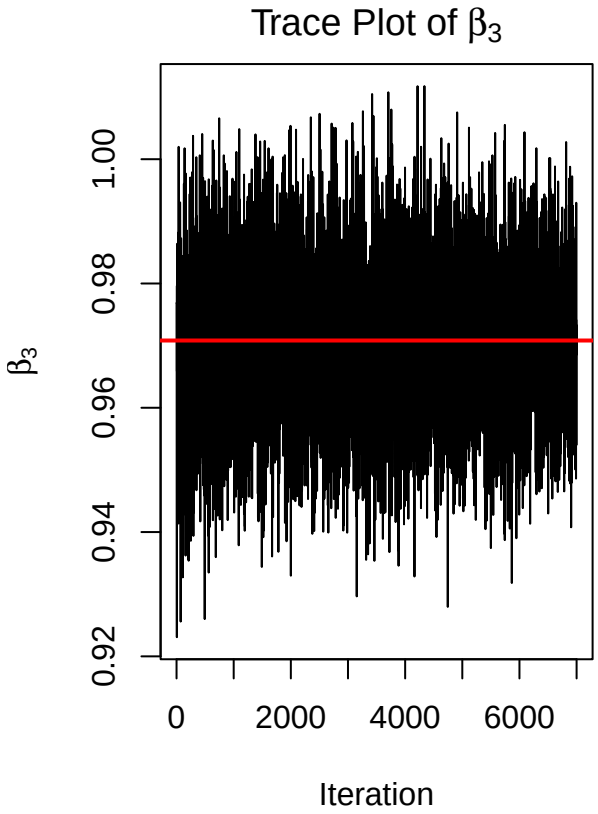
where m_j and $s^{2(s)}$ are the sample mean and variance of the beta accepted until iteration s .

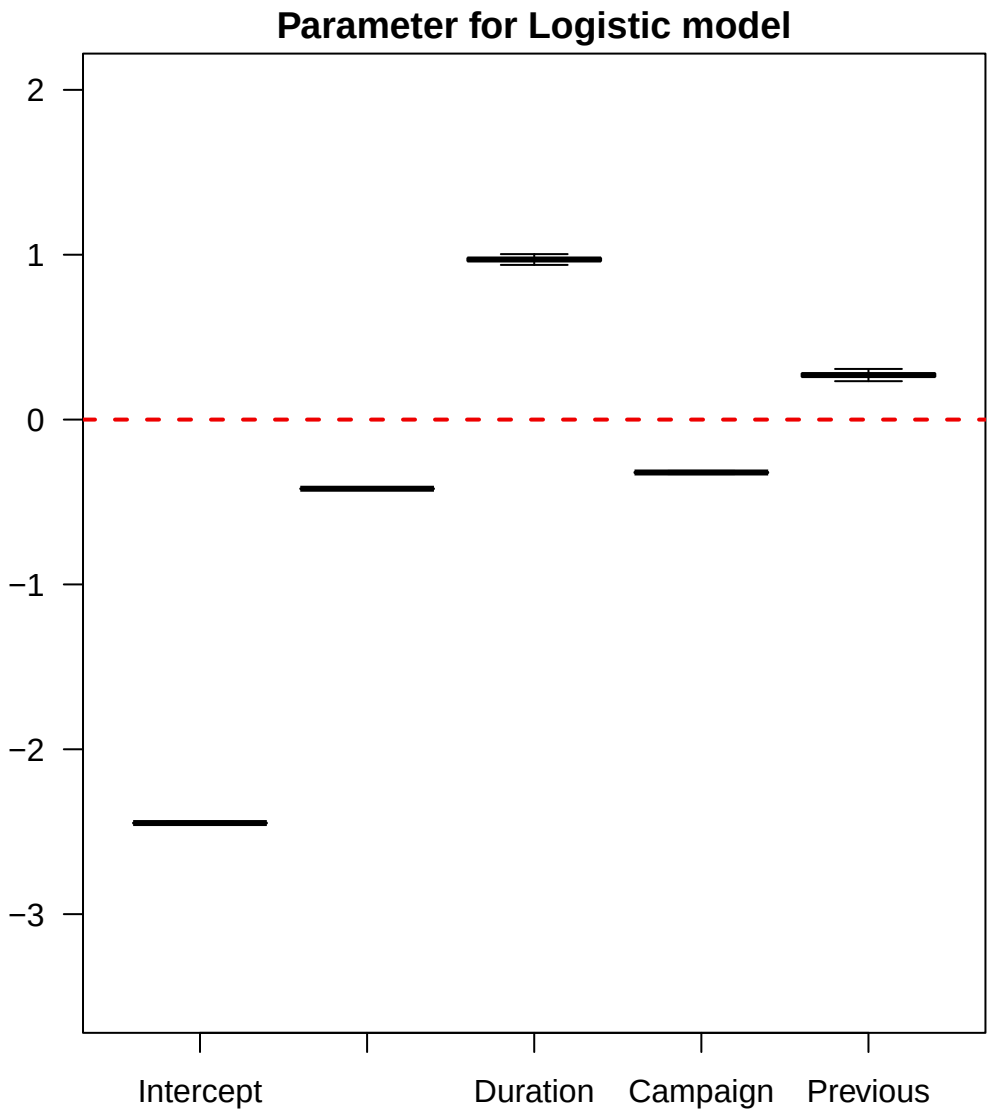
Implementation

The sequences of samples obtained from the algorithm represents the approximation of the posterior distribution for each parameter. The different empirical distributions are helpful for updating our uncertainty around the parameter estimates.

```
##                    Mean         Q5        Q95
## Intercept    -2.4473941 -2.4523730 -2.4425061
## Housing|yes  -0.4194386 -0.4226092 -0.4162742
## Duration      0.9708230  0.9503081  0.9909708
## Campaign     -0.3210029 -0.3271369 -0.3149283
## Previous      0.2699672  0.2478424  0.2921569
```



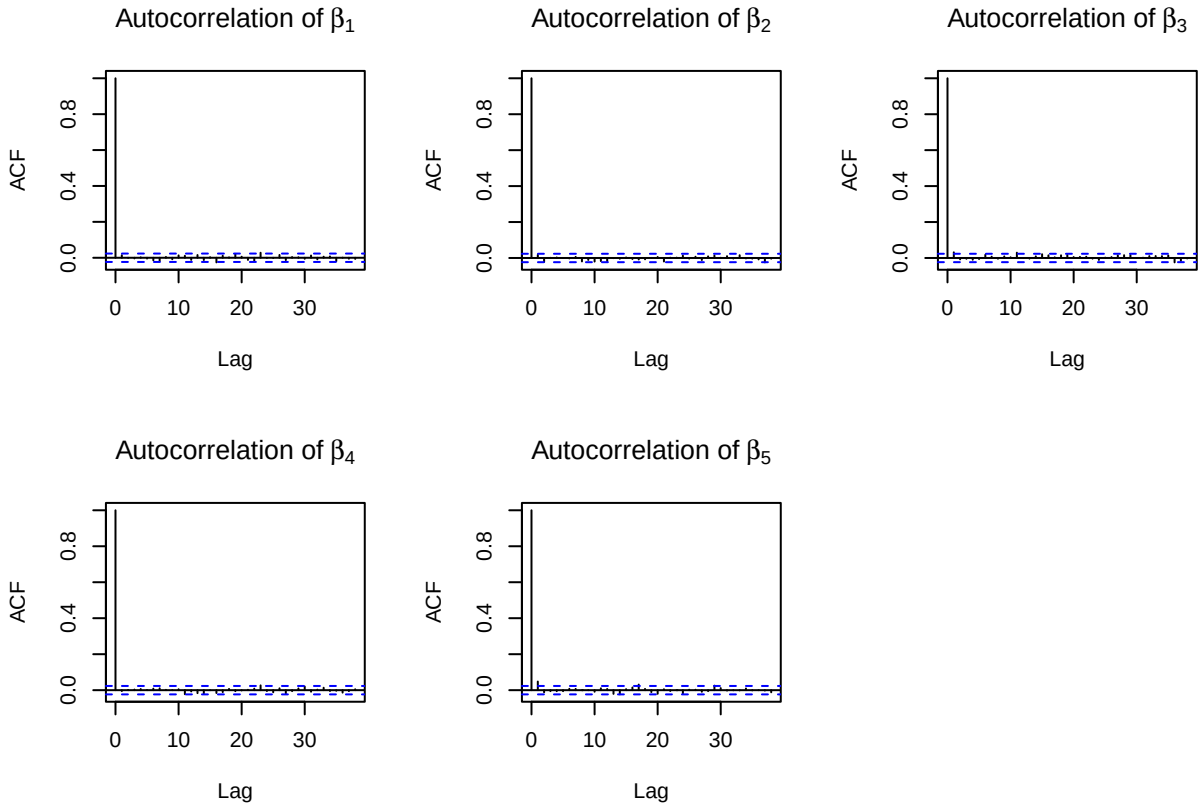




All coefficients in the model appear to be significantly different from zero. This suggests that each predictor included in the model plays a statistically significant role in explaining the variation in the binary response variable. The signs of the coefficients show that already existing house loans and a number of contacts above average have a potential deterrent effect on client subscription (negative coefficients). Conversely, longer calls and more frequent contacts before the campaign, having positive coefficients, have a positive influence on subscription likelihood.

Diagnostics

To assess the accuracy of the approximation provided by the Metropolis-Hastings algorithm, we perform diagnostics using tools such as autocorrelation, Geweke test and effective sample size.



The autocorrelation plots for the model parameters reveal negligible dependencies within the chain, as all autocorrelation function (ACF) values are very low. This suggests that successive samples from the MCMC simulations are almost independent and there aren't trends or cycle, indicating no significant autocorrelation issues that would necessitate thinning.

Geweke Test

We perform a Geweke test, which compares statistics from different portions of the chain, to verify whether the chain is stationary. The initial and final $\alpha = 10\%$ quantiles of the posterior distribution of the parameters will be used.

```
##
## Fraction in 1st window = 0.1
## Fraction in 2nd window = 0.1
##
##   Intercept Housing|yes    Duration    Campaign    Previous
##      1.5035      1.2084     -3.6312      1.2006      0.9497
```

The results highlight that *Duration* may have problems of convergence due to the difference between the value in the α initial quantile of the samples and the one in the final.

Effective Sample Size

```r
effectiveSize(x = beta.post_sample.new)
```

```
##   Intercept Housing|yes    Duration    Campaign    Previous
##    7000.000    7073.618    6571.056    7000.000    6357.940
```

$$ESS = \frac{G}{1 + 2\sum_{g=1}^{G} acf_g}$$

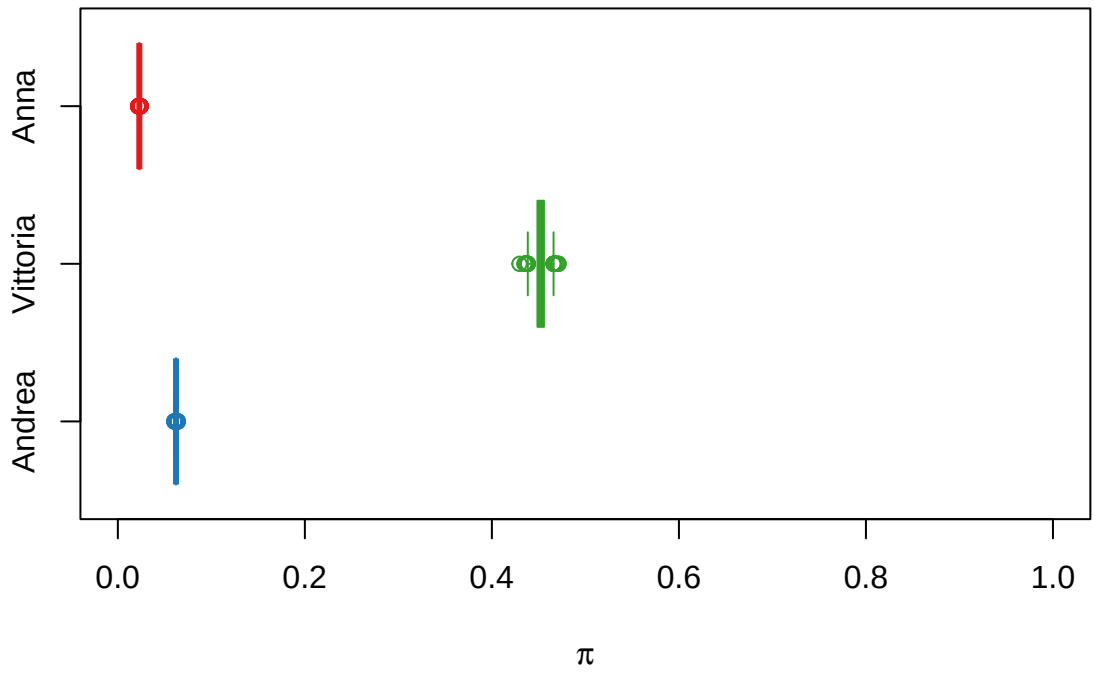
The ESS represents a measure of the number of effectively independent draws obtained from the algorithm, compared to the number of draws (we simulated 7000 iteration through the Metropolis-Hastings Algorithm). Discrepancies are due to correlation in the sequence, it decrease as the autocorrelation increase.

Predictions

We predict the probability of success of the campaign for 3 individuals that present different characteristics:

```r
x.star.andrea = c(1, 0, 150, 5, 0)
x.star.vitto = c(1, 0, 600, 0, 2)
x.star.anna = c(1, 0, 10, 10, 0)
```

```
##       Name Probability
## 1   Andrea  0.06220139
## 2 Vittoria  0.45220429
## 3     Anna  0.02297304
```



The second subject, having higher values for call time and previous contacts, which were the variables that influence positively the response the most, has the highest probability of success. The third subject, with the lowest phone call time and highest contact record, presents an extremely low probability.

Comparison with the Probit model

We now compare the previous model with a second one obtained through probit regression. We combine the prior and likelihood, as already presented, to obtain the posterior distribution using a different link function:

$$P(Y = 1 \mid X) = \Phi(\beta^T X)$$

The link function is now the cumulative distribution function of a standard Normal random variable, as its values range between $(0, 1)$ The addition of a latent variable z_i such that for every y_i:

$$z_i|\beta \sim N(\beta^\top x_i, 1)$$

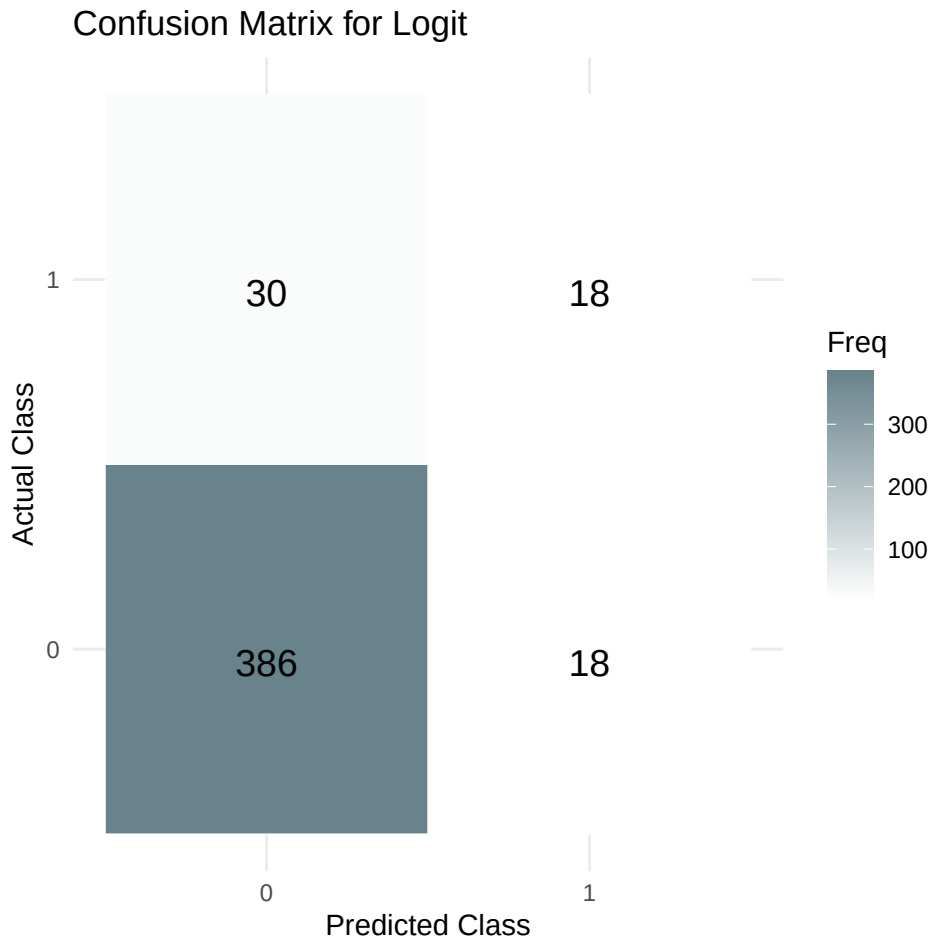
allows to obtain full conditional distributions in closed form, and therefore to implement a Gibbs sampler.

```
##               Probit.b    Logit.b
## Intercept   -1.4072922 -2.4473941
## Housing|yes -0.2163029 -0.4194386
## Duration     0.5252195  0.9708230
## Campaign    -0.1830073 -0.3210029
## Previous     0.1507071  0.2699672
```

The estimates obtained for the coefficients of the probit model show slight differences compared to those from the logit model. However, all coefficients are coherent in sign, indicating consistent directionality of the effects of predictors on the response variable across both models. This consistency supports the robustness of our findings despite the choice of link function.

```
##       Name  MeanLogit MeanProbit
## 1   Andrea 0.06220139 0.05876772
## 2 Vittoria 0.45220429 0.42588647
## 3     Anna 0.02297304 0.01685437
```

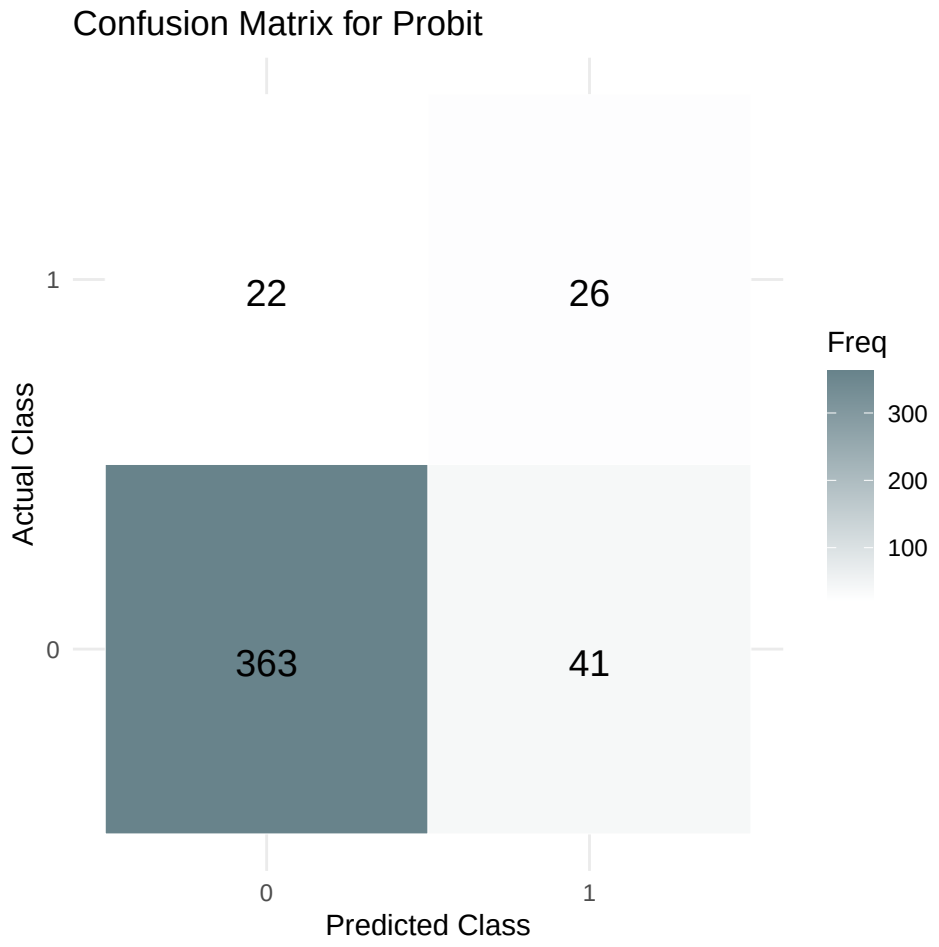
Additionally, the predictions for the three subjects already presented in the project are very similar, regardless of whether they come from the probit or logit model.




```
## Accuracy: 0.8938053

## Specificity: 0.9554455

## Sensitivity: 0.375
```




```
## Accuracy: 0.8606195
```

```
## Specificity: 0.8985149
```

```
## Sensitivity: 0.5416667
```

To assess the predictive performance of the models and compare them, we used confusion matrices with a threshold of 0.3, determined through several trials. The accuracy (overall correct classification) was good for both models, with the logit model achieving 89% accuracy and the probit model 86%. The specificity (proportion of negatives correctly classified) was also high, at 95% for the logit model and 89% for the probit model. However, sensitivity (proportion of positives correctly classified) was less satisfactory, particularly for the logit model, which achieved only 37%. On the other hand, the overall number of wrong predictions was higher in the probit model.

Conclusion

In conclusion, the models we implemented consist of a collection of significant predictors that have a remarkable influence on the response variable. The results from the regression using both the logit and probit models are largely consistent and do not differ significantly. However, when these models are applied to predict the outcome of the campaign, they exhibit some issues—either with false negatives or with overall wrong predictions. This indicates that while our models are robust in identifying influential factors, there is room for improvement in their predictive accuracy.