STATISTICS IN TRANSITION new series, December 2016
Vol. 17, No. 4, pp. 763–780

INFORMATIVE VERSUS NON-INFORMATIVE PRIOR DISTRIBUTIONS AND THEIR IMPACT ON THE ACCURACY OF BAYESIAN INFERENCE

Wioletta Grzenda[1]

ABSTRACT

In this study the benefits arising from the use of the Bayesian approach to predictive modelling will be outlined and exemplified by a linear regression model and a logistic regression model. The impact of informative and non-informative prior on model accuracy will be examined and compared. The data from the Central Statistical Office of Poland describing unemployment in individual districts in Poland will be used. Markov Chain Monte Carlo methods (MCMC) will be employed in modelling.

Key words: Bayesian approach, regression models, a priori information, MCMC.

1. Introduction

For data mining techniques, classification and regression methods play an important role. The choice of an appropriate model is the basis of data analyses. The key advantage of the Bayesian approach is the ability to include additional information that is external to the sample in the modelling process (Lancaster, 2004). In Bayesian analysis, statistical inference is based on posterior distributions, which combine prior information with sample-based information. The impact of prior information on estimation model parameters in the parametric survival models has been investigated in (Grzenda, 2013), among others. In modelling, taking into account prior information has also an influence on the predictive power of a model.

The Bayesian model selection criteria frequently correspond to finding a model, which is characterised by a maximum posterior probability while considering model selection in the context of decision problems. The primary objective of this paper is to analyse the impact of prior information on the predictive power of a model using selected measures assessing the accuracy of prediction. Particular attention is paid to the selection of informative versus non-

[1] Warsaw School of Economics, Collegium of Economic Analysis, Institute of Statistics and Demography. E-mail: wgrzend@sgh.waw.pl.

informative prior distribution. This is because the appropriate selection of a priori distribution may result in more accurate models.

Moreover, in this paper, the impact of informative and non-informative prior distributions on the accuracy of both classification and regression have been investigated. What should be emphasised in this context is that in Bayesian methods (Congdon, 2006; Gelman et al., 2000) parameters of a model are treated as random variables. Let θ denote the estimated parameter, and $\mathbf{x}$ observed data. The initial knowledge about the parameter θ is represented by prior distribution $p(\theta)$. The Bayesian inference approach is based on posterior distribution, which is determined in the following way (Bolstad, 2007):

$$p(\theta \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \theta)p(\theta)}{\int p(\mathbf{x} \mid \theta)p(\theta)d\theta}.$$

This equation is expressed in an equivalent proportional form:

$$p(\theta \mid \mathbf{x}) \propto p(\mathbf{x} \mid \theta)p(\theta).$$

In Bayesian approach, posterior distribution includes all available knowledge about the unknown parameter. This is prior information and information derived from data. The posterior distribution can be summarized by one statistic. Most frequently, this is the posterior mean as it minimizes a posterior mean square error. It is given by the formula:

$$E(\theta \mid \mathbf{x}) = \int \theta p(\theta \mid \mathbf{x})d\theta.$$

Frequently, instead of a single parameter θ, the parameter vector $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_k]^T$ is considered. The inference about any element of vector $\boldsymbol{\theta}$ is performed using marginal distribution, which is obtained by integrating the joint posterior distribution over the remaining coordinates. Given the complexity of calculations, the Markov chain Monte Carlo methods (MCMC) are used in practice (Congdon, 2006). The most famous algorithm among these methods is the Metropolis algorithm. In this paper the adaptive rejection Metropolis sampling algorithm (ARMS), which is a generalization of the Metropolis algorithm, has been used.

In this study, multiple regression models and logistic regression models have been estimated with informative and non-informative prior distributions. Based on obtained posterior distributions of model parameters, the posterior means have been calculated. These posterior means have been used as estimates for unknown parameters of the model (Lancaster, 2004). Next, the selected measures for model accuracy have been determined and compared. Many statistics can be used to measure the accuracy of models (Japkowicz and Shah, 2011; Provost and Fawcett, 2013). In the case of classification, the key measures are the incorrect

classification rate and confusion matrix, while in the case of regression the mean square error, median square error and maximum absolute error are frequently used. The predictive power of competing models can be compared based on Lift and ROC curves.

2. The scope of research

In this paper data from the Central Statistical Office of Poland, describing the districts in Poland, has been used. The object of this study is unemployment rate in districts in Poland in the year 2014. The characteristics of posterior distribution obtained for the previous year have been suggested as prior information for modelling data for the next year. Therefore, two sets of data have been created to model the unemployment rate in two successive years: 2013 and 2014. The number of observations in both data sets is the same, namely 380.

The examined feature is the unemployment rate in districts in Poland; in August 2013 the average was 15.92%, whereas in August of the next year the average was 14.32%. Moreover, for the purpose of this research, i.e. the investigation of classification accuracy, a binary variable *unemployment* has been created based on the continuous variable *unemployment_rate*. The variable *unemployment* differentiates the districts into those with low unemployment below 10% and the remaining ones. In 2013 there were 61 (16.05%) districts with the unemployment rate below 10%, whereas in 2014 there were 93 districts (24.47%). The unemployment may be defined and explored in many ways, but it is worth emphasising that a significant spatial diversity of the unemployment rate is observed in Poland (Gołata, 2004). The subject matter of this study is registered unemployment, including the unemployed registered in the district labour offices and seeking employment through these offices.

The preliminary data analysis including variable significance assessment, model adjustment to fit the observed data, model correctness verification and predictive power assessment (Lancaster, 2004) reduced the initially proposed set of variables to the following variables:

- salary – the amount of average monthly gross wages and salaries in thousand zlotys (mean=3.42, min=2.54, max=6.81);
- number_children – the number of children aged 3-5 per one place in nursery school (mean=1.48, min=0.77, max=5.11);
- flats – the number of flats ready for occupancy per 1000 residents (mean=2.92, min=0.16, max=15.26);
- EU_funds – the total value of contracts signed for financing in million zlotys per 1000 residents (mean=10.5, min=1.83, max=107.74);
- farm – the average area of individual farm (farms over 1 hectare have been investigated): 1 - less than 10 hectares (49.21%), 2 - from 10 to 15 hectares (30.79%), 3 - 15 hectares and over (20%);

- innovation – the average share of innovative companies in the total number of companies in %: 1 - less than 13% (35.26%), 2 - from 13 to 15% (49.21%), 3 - 15% and over (15.53%).

Variable characteristics for 2014 have been given in parentheses. The characteristics of the districts such as education, road infrastructure or population density, which were investigated in other studies such as (Gołata, 2004) have turned out to be statistically insignificant.

3. The multiple regression models

3.1. Bayesian multiple regression model

Let $\mathbf{y} = [y_1, \ldots, y_n]^T$ be the vector of observed values of the dependent variable $\mathbf{X}$, $(n \times k)$ be a matrix of independent variables, and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]^T$ be a vector of regression coefficients. The classical linear regression model can be expressed as follows (Draper and Smith, 1981):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ denotes an error vector, $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I})$.

In Bayesian approach (Gelman et al., 2000; Gill, 2008), the regression coefficients $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]^T$ are random variables. Let $p(\boldsymbol{\beta})$ denote their joint prior distribution and let us assume that the elements of vector $\boldsymbol{\beta}$ are independent. Then, we have the following the likelihood function:

$$L(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}, \mathbf{y}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right].$$

Then, based on Bayes' theorem, posterior distribution is given by:

$$p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}, \mathbf{y}) \propto L(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}, \mathbf{y}) p(\boldsymbol{\beta}) p(\sigma^2).$$

For model parameters, various prior distributions can be selected. The Bayesian approach with an informative prior allows us to incorporate additional information. If we do not have such information, then a non-informative prior can be selected. For regression coefficients $\boldsymbol{\beta}$, the most frequent normal prior distribution is selected: $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta)$. Assuming that the average equals 0 and there is a suitably small variation, a non-informative prior distribution is obtained: $\boldsymbol{\beta} \sim N(\mathbf{0}, 10^6 \mathbf{I})$. For the parameter σ^2, inverse gamma distribution is selected most frequently.

3.2. Key accuracy indicators of multiple regression models

While examining the accuracy of the model it is important to determine how the estimated values differ from actual values present in the training data. There are many ways to calculate the error, i.e. the difference between the estimated and actual values. The most natural one (Provost and Fawcett, 2013) is determining the absolute error:

$$AE = \left| y_i - \hat{y}_i \right| .$$

The maximum absolute error is a useful measure of prediction accuracy in the case of extreme values:

$$MAE = \max_i \left| y_i - \hat{y}_i \right| .$$

The sum of the squares error is a commonly used criterion for model accuracy. It is a natural consequence of estimating parameters of the classical regression model using least squares methods. This measure expresses the total value of the estimate error when the regression equation is used:

$$SSE = \sum_i \left(y_i - \hat{y}_i \right)^2 .$$

The degree of regression fit as an approximate linear relationship between the dependent variable and the explanatory is given by the coefficient of determination:

$$R^2 = \frac{SSR}{SST},$$

where $SSR = \sum_i \left(\hat{y}_i - \bar{y} \right)^2$ denotes the sum of squares regression and SST=SSR+SSE the sum of squares total, respectively. Finally, the mean squared error is defined as:

$$MSE = \frac{SSE}{n - m - 1},$$

where n is the number of observations, and m is the number of explanatory variables.

3.3. The specification and estimation of Bayesian multiple regression models

In this section, the multiple regression models with informative and non-informative prior distributions are discussed. In the first model developed for data from 2014, a priori distribution that has a minimal impact on posterior distribution has been used for all model parameters (Gelman et al., 2000). Therefore, non-informative independent normal prior distributions with mean equalling 0 and

variance 10^6 for regression coefficients, as well as inverse gamma distribution for the parameter σ^2, have been used. In all investigated models, the number of burn-in samples is assumed to be 2000 and posterior samples equals 10000 in order to minimize the effect of initial values on posterior inference. The highest posterior density (HPD) intervals for all parameters in all models have been determined for α=0.05. The characteristics of prior distributions and posterior distributions of the first model parameters for data from 2014 are presented in Table 1.

Table 1. The prior and posterior distributions

Parameter	Model 1					
	Prior distributions		Posterior distributions			
	Mean	Standard dev.	Mean	Standard dev.	HPD	
Intercept	0	10^6	25.7691	2.5752	20.4681	30.5697
Salary	0	10^6	-2.8037	0.5352	-3.8700	-1.7685
Number_children	0	10^6	3.3589	0.4622	2.4530	4.2562
EU_funds	0	10^6	-0.1090	0.0286	-0.1643	-0.0521
Farm1	0	10^6	-3.7221	0.6291	-4.9466	-2.4826
Farm2	0	10^6	-5.2634	0.9361	-7.1080	-3.4612
Innovation1	0	10^6	-2.1867	0.8343	-3.8149	-0.5576
Innovation2	0	10^6	-2.9499	1.0464	-5.0837	-0.9786
Dispersion	IG		21.1553	1.5442	18.2595	24.2207

Based on the highest posterior density intervals (Bolstad, 2007), all variables are statistically significant for α=0.05. The convergence of generated Markov chain has been verified by several tests and graphically. The result of Geweke's test (Geweke, 1992) is included in Table 2. The graphs for generated chains are presented in the Figures 1-9. The results show no indication that the Markov chain has not converged for all the parameters of the investigated model at any significant level. Moreover, the Monte Carlo standard error (MCSE) is presented.

Table 2. Geweke convergence diagnostics and MCSE

Parameter	Model 1		
	Geweke diagnostics		MCSE
	z	p-value	
Intercept	1.6547	0.0980	0.0969
Salary	-0.8009	0.4232	0.0078
Number_children	-0.7595	0.4475	0.0052
EU_funds	-0.9591	0.3375	0.0003
Farm1	-0.3756	0.7072	0.0176
Farm2	-1.4492	0.1473	0.0469
Innovation1	-1.9345	0.0531	0.0399
Innovation2	-1.7811	0.0749	0.0568
Dispersion	1.8983	0.0577	0.0162

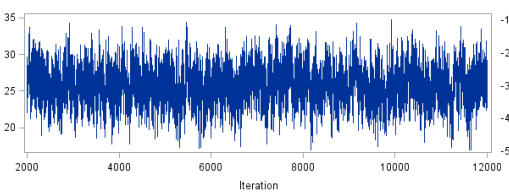

Figure 1. Trace Plots for *Inercept*

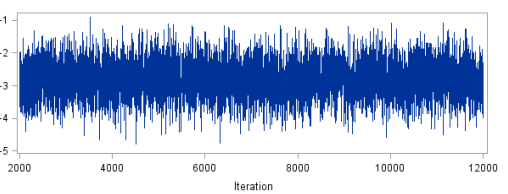

Figure 2. Trace Plots for *Salary*

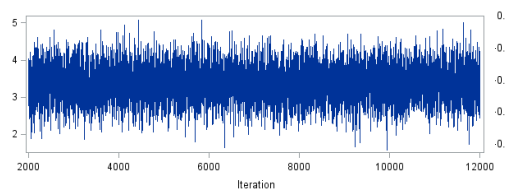

Figure 3. Trace Plots for *Number_children*

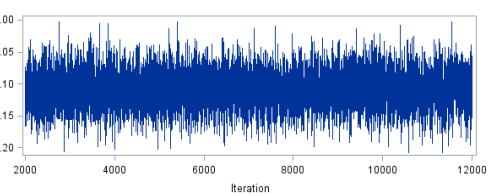

Figure 4. Trace Plots for *EU_funds*

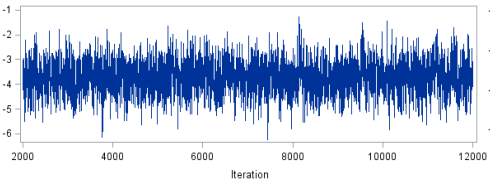

Figure 5. Trace Plots for *Farm1*

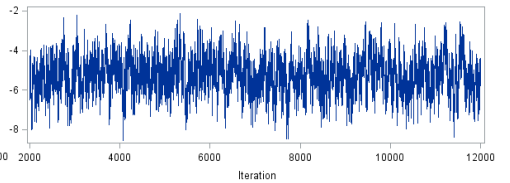

Figure 6. Trace Plots for *Farm2*

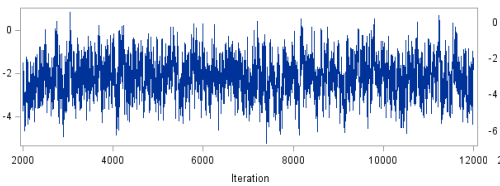

Figure 7. Trace Plots for *Innovation1*

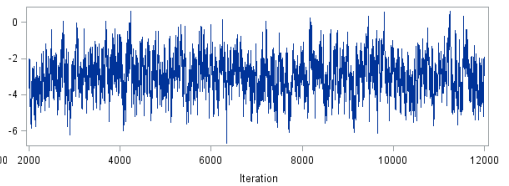

Figure 8. Trace Plots for *Innovation2*

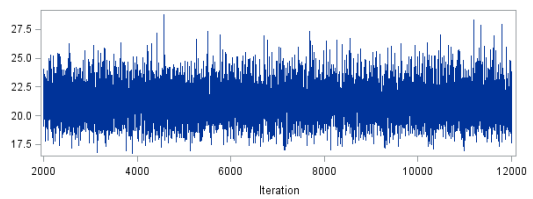

Figure 9. Trace Plots for *Dispersion*

Next, the multiple regression models for data from 2013 have been estimated. In order to obtain objectively correct results, non-informative prior distributions have been assumed in the model for data from 2013, as in the previous model. The obtained characteristics of posterior samples for data from 2013 have been used as prior information for the regression coefficients in the model for data from 2014. The prior and posterior distributions for second model parameters for data from 2014 are given in Table 3. All variables are again statistically significant for α=0.05.

Table 3. The prior and posterior distributions

Parameter	Model 2					
	Prior distributions		Posterior distributions			
	Mean	Standard dev.	Mean	Standard dev.	HPD	
Intercept	26.8856	2.3281	25.5936	1.2719	23.2001	28.2022
Salary	-3.0471	0.5471	-2.9138	0.3281	-3.5662	-2.2840
Number_children	3.2935	0.4026	3.3003	0.2926	2.7320	3.8825
EU_funds	-0.1058	0.0283	-0.1099	0.0197	-0.1479	-0.0711
Farm1	-5.6123	0.6687	-4.5617	0.4342	-5.4003	-3.7021
Farm2	-4.6628	0.7196	-4.8026	0.4872	-5.7637	-3.8706
Innovation1	-2.7432	0.7040	-1.7124	0.4616	-2.6566	-0.8517
Innovation2	-0.3766	0.5933	-1.3999	0.4505	-2.2689	-0.5033
Dispersion	IG		21.2691	1.5815	18.2931	24.4650

The result of Geweke's test (Geweke, 1992) and the Monte Carlo standard error are included in Table 4. The results show no indication that the Markov chain has not converged for all the parameters of the investigated model at the significance level 0.01. The values of Monte Carlo standard errors for all model 2 parameters have been lower than in model 1.

Table 4. Geweke convergence diagnostics and MCSE

Parameter	Model 2		
	Geweke diagnostics		MCSE
	z	p-value	
Intercept	0.5981	0.5498	0.0154
Salary	0.1330	0.8942	0.0034
Number_children	-0.6970	0.4858	0.0029
EU_funds	-0.2500	0.8026	0.0002
Farm1	-2.1693	0.0301	0.0078
Farm2	-0.7126	0.4761	0.0095
Innovation1	0.4638	0.6428	0.0087
Innovation2	0.3302	0.7412	0.0096
Dispersion	1.6572	0.0975	0.0161

For both models, a deviance information criterion (DIC) has been calculated. For the first model DIC is 2245.575, for the second one it is 2244.557, the results do not differ significantly.

The estimated posterior means have been selected as estimation for the unknown model parameters for both models. With these assumptions, selected measures of model accuracy have been calculated (Table 5).

Table 5. Precision and accuracy of models

Statistics		**Model 1 Non-informative Priors**	**Model 2 Informative Priors**
Maximum Absolute Error	*MAE*	21.621	20.816
Sum of Squares Error	*SSE*	16242.310	15487.660
Sum of Squares Regression	*SSR*	13205.760	12639.100
Coefficient of Determination	R^2	0.448	0.449
Mean Squared Error	*MSE*	43.662	41.632

The obtained results indicate that the model with informative prior is the one with greater prediction accuracy. Moreover, these results indicate that about 45% of *unemployment_rate* variable variance is explained by the estimated models.

The estimated values of multiple regression models show that only the variable describing the number of children aged 3-5 per one place in nursery school has a positive impact on unemployment in the analysed districts. Thus, the greater the number of children per one place in nursery school, the higher the unemployment levels. The results also indicate that the lower salaries in districts, the higher unemployment. The unemployment rate in a given district also depends on the total value of EU contracts signed for financing – the smaller the value of grants, the higher unemployment. Moreover, the study found that the more fragmented farms and the lower the average share of innovative companies in the total number of companies, the higher the unemployment rates.

4. The logistic regression models

4.1. Bayesian logistic regression model

The logistic regression models (Finney, 1972; Hosmer and Lemeshow, 2000) are very often used in the study of socio-economic phenomena when a binary dependent variable is considered. These models are also applied to estimate the probability of belonging to a given class in classification tasks (Japkowicz and Shah, 2011).

Let us consider a dependent variable that takes only two values. Let $y_i = 1$ indicate the presence, and $y_i = 0$ the absence of the event, for $i = 1,\ldots,n$. Moreover, let p_i denote the probability that $y_i = 1$, $p_i = P(y_i = 1)$. Let $\mathbf{x}_i = [1, x_{i1},\ldots, x_{ik}]^T$ be a vector of independent variables, and $\boldsymbol{\beta} = [\beta_0, \beta_1,\ldots,\beta_k]$ be a vector of regression coefficients. Let $\text{logit}(p_i) = \boldsymbol{\beta}\mathbf{x}_i$, then the classical logistic regression model can be expressed as follows:

$$p_i = \frac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}\mathbf{x}_i)}.$$

The likelihood function over a data set for *n* subjects is:

$$L(\boldsymbol{\beta} \mid \mathbf{y}) = \prod_{i=1}^{n}\left[(p_i)^{y_i}(1-p_i)^{(1-y_i)}\right] = \prod_{i=1}^{n}\left[\left(\frac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}\mathbf{x}_i)}\right)^{y_i}\left(1-\frac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}\mathbf{x}_i)}\right)^{(1-y_i)}\right]$$

In this paper, Bayesian logistic regression models are investigated (Albert and Chib, 1993; Congdon, 2006; Gelman et al., 2000). Assuming normal prior distribution $\beta_j \sim N(\mu_j, \sigma_j^2)$ for regression coefficients, and each of them being independent from the other, the posterior distribution is given by:

$$p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}) \propto \prod_{i=1}^{n}\left[\left(\frac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}\mathbf{x}_i)}\right)^{y_i}\left(1-\frac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}\mathbf{x}_i)}\right)^{(1-y_i)}\right]$$

$$\cdot \prod_{j=0}^{k}\left[\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left\{-\frac{1}{2}\left(\frac{\beta_j-\mu_j}{\sigma_j}\right)^2\right\}\right].$$

4.2. Key accuracy indicators of logistic regression models

The evaluation of the accuracy of a logistic regression model can be performed in many ways (Hosmer and Lebeshow, 2000). If the purpose of the modelling is to obtain the best possible classification (Provost and Fawcett, 2013), the following measures are the most common: the confusion matrix or classification table, the accuracy rate or interchangeably misclassification error rate. Graphically, the classification accuracy can be verified with ROC curve and LIFT curve (Japkowicz and Shah, 2011). Models with good classification capacity should be characterized by a high accuracy and a low rate of

misclassification. The results for the logistic regression model can be summarized in a classification table:

		Observed	
		POSITIVE	NEGATIVE
Predicted	YES	True positive	False positive
	NO	False negative	True negative

The basic measure for assessing the accuracy of the model in terms of classifying individual observations into the groups designated by the dependent variable is the accuracy of classification, i.e. the percentage of correct decisions:

$$Accuracy = \frac{Number\ of\ \ correct\ decisions\ made}{Total\ \ number\ of\ \ decisions\ made}$$

Alternatively, the misclassification error rate is calculated:

$$MISC = 1 - Accuracy$$

Based on the table such measures as sensitivity or true positive rate (TPR) and specificity or true negative rate (TNR) are often calculated:

$$TPR = \frac{true\ \ positive}{true\ \ positive + \ false\ negative},$$

$$TNR = \frac{true\ negative}{true\ negative + \ false\ \ positive}.$$

To determine the ROC curve, FPR (false positive rate) is calculated as 1-TNR. The ROC curve is formed by presenting FPR values on the axis *X*, and TPR values on the axis *Y*.

Model adjustment in terms of data and the prognostic effectiveness of competing models can also be compared using the LIFT curve. For a given model the LIFT curve compares the predictive model to no model (pick randomly):

$$\frac{True\ \ positive\ of\ \ Model}{True\ \ positive\ of\ no\ \ Model}.$$

4.3. The specification and estimation of Bayesian logistic regression models

Similarly to multiple regression models, Bayesian logistic regression models with non-informative and informative prior distributions were compared. The general assumptions regarding Bayesian estimation for logistic regression models were the same as in the case of multiple regression models. A model for the data from the year 2014 has been estimated, using non-informative normal prior distribution for all regression coefficients (Model 3). In Table 6, prior distribution settings and posterior distribution statistics for Model 3 are shown. For α=0.05, all variables are statistically significant except one level of *Farm* variable.

Table 6. The prior and posterior distributions

Parameter	Model 3					
	Prior distributions		Posterior distributions			
	Mean	Standard dev.	Mean	Standard dev.	HPD	
Intercept	0	10^6	-6.9112	1.7665	-10.5488	-3.6320
Salary	0	10^6	1.2748	0.3379	0.6280	1.9384
Number_children	0	10^6	-1.9011	0.4606	-2.8083	-1.0316
Flats	0	10^6	0.2332	0.0856	0.0652	0.4013
EU_funds	0	10^6	0.0535	0.0196	0.0170	0.0934
Innovation1	0	10^6	1.4576	0.4885	0.5292	2.4134
Innovation2	0	10^6	2.3524	0.6803	1.0295	3.6720
Farm1	0	10^6	0.2196	0.4166	-0.5637	1.0672
Farm2	0	10^6	2.3576	0.6328	1.0999	3.5671

In Table 7, the results of Geweke's test (Geweke,1992) and the values of Monte Carlo standard error are shown. The study failed to reject the null hypothesis that the chains generated for individual model parameters converge at any level of significance. The figures depicting generated chains confirmed the inference regarding the convergence of these chains.

Table 7. Geweke convergence diagnostics and MCSE

Parameter	Model 3		
	Geweke diagnostics		MCSE
	z	p-value	
Intercept	1.8952	0.0581	0.0764
Salary	-1.2440	0.2135	0.0060
Number_children	-0.0992	0.9210	0.0067
Flats	-0.6526	0.5140	0.0011
EU_funds	-1.4819	0.1384	0.0002
Innovation1	-1.8530	0.0639	0.0244
Innovation2	-1.7862	0.0741	0.0424
Farm1	-1.0598	0.2892	0.0142
Farm2	-1.5770	0.1148	0.0381

Next, Bayesian logistic regression model has been estimated using informative prior distributions. The estimation was performed in the same way as for multiple regression models. The model is hereinafter referred to as Model 4. The results of the model estimation are provided in Table 8.

Table 8. The prior and posterior distributions

Parameter	Model 4					
	Prior distributions		Posterior distributions			
	Mean	Standard dev.	Mean	Standard dev.	HPD	
Intercept	-5.3563	1.8043	-5.4033	0.9400	-7.2131	-3.5243
Salary	1.0686	0.3208	1.0654	0.2048	0.6661	1.4633
Number_children	-2.9115	0.6143	-2.2641	0.3597	-2.9848	-1.5786
Flats	0.3105	0.0949	0.2551	0.0629	0.1367	0.3831
EU_funds	0.0516	0.0198	0.0520	0.0138	0.0244	0.0782
Innovation1	1.8938	0.5578	1.3484	0.3230	0.7076	1.9705
Innovation2	1.2809	0.5354	1.7586	0.3657	1.0718	2.5072
Farm1	1.0792	0.6046	0.4214	0.3319	-0.2251	1.0632
Farm2	1.7633	0.6238	2.0049	0.3765	1.2680	2.7459

Table 9 provides the results of Geweke's test and MCSE values for Model 4. The results show that the study failed to reject the null hypothesis that the chains generated for individual model parameters converge at any level of significance. The MCSE values for all parameters of Model 4 are lower than the corresponding values for Model 3.

Table 9. Geweke convergence diagnostics and MCSE

Parameter	Model 3		
	Geweke diagnostics		MCSE
	z	p-value	
Intercept	1.8952	0.0581	0.0139
Salary	-1.2440	0.2135	0.0021
Number_children	-0.0992	0.9210	0.0049
Flats	-0.6526	0.5140	0.0007
EU_funds	-1.4819	0.1384	0.0001
Innovation1	-1.8530	0.0639	0.0076
Innovation2	-1.7862	0.0741	0.0111
Farm1	-1.0598	0.2892	0.0078
Farm2	-1.5770	0.1148	0.0107

For the third model, the DIC value equals 336.117, while the value of the same indicator for Model 4 is lower and equals 331.776. Therefore, Model 4 is a better model out of the two models.

The average values of the posterior distributions of the third and fourth model were used as the estimation for unknown model parameters. Next, the performance indicators for both models were calculated and compared. The lower misclassification error rate observed in the case of the model with informative prior distribution indicates that it is a model of higher accuracy (Table 10).

Table 10. Geweke convergence diagnostics and MCSE

Statistics		Model 3 Noninformative Priors	Model 4 Informative Priors
Accuracy Rate	*AR*	67.11	70.53
Misclassification Error Rate	*MICS*	32.89	29.47
True Positive Rate	*TPR*	0.495	0.516
True Negative Rate	*TNR*	0.728	0.766

ROC curves for models with informative and non-informative prior distributions are provided in Figure 10. This confirms that the model based on informative prior distribution is a model with better classification properties.

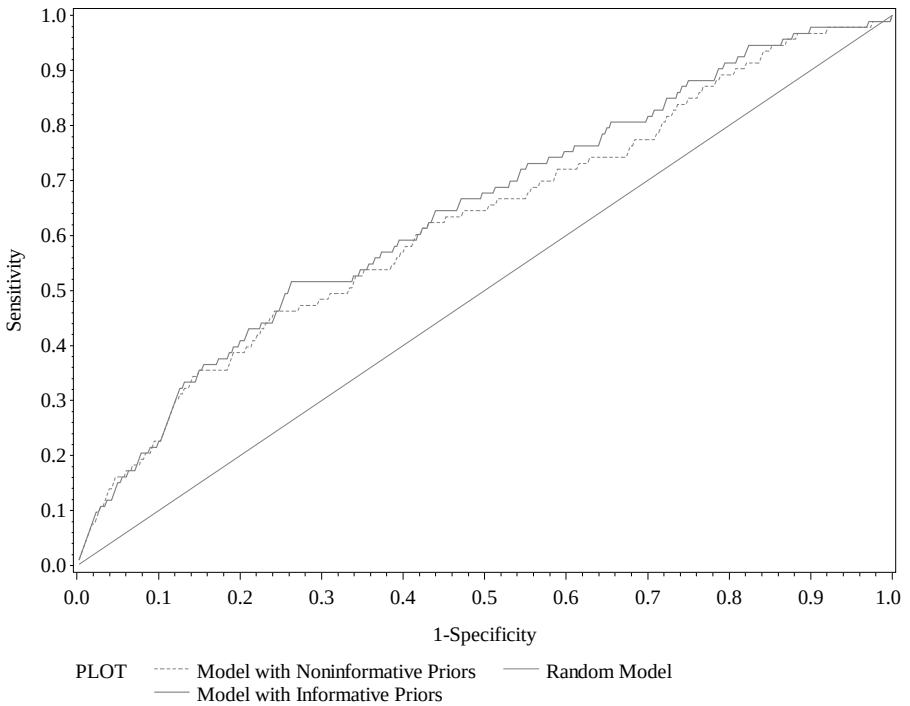

Figure 10. The ROC curve

LIFT curves for the model with non-informative and the model with informative prior distributions are shown in Fig. 11. The LIFT trend indicates that a model matches well the data (Tufféry, 2011). For every decile, the LIFT curve developed for the model based on informative prior distributions is located above the curve formed for a model created with non-informative prior distributions. Therefore, the model using informative prior distributions demonstrates better classification capabilities.

To sum up, all the analysed accuracy indicators show that the logistic regression model with informative prior distributions yields better classification capabilities.

Moreover, the estimation of logistic regression parameters shows that larger values of all the variables except for *number children* increase the probability of the unemployment rate in the district being below 10%. The higher the salaries, the bigger the number of flats ready for occupancy, and the larger the EU funds, the higher the chances of a low unemployment rate in the district. Moreover, the less fragmented farms and the bigger proportion of innovative enterprises, the higher the probability of the unemployment level in the district being below 10%.

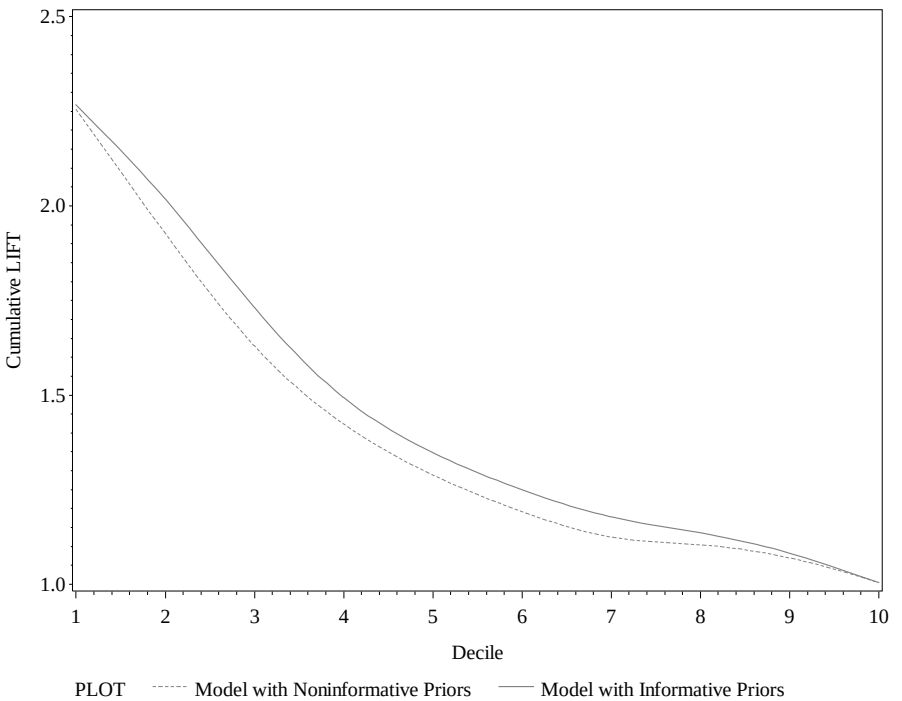

Figure 11. The LIFT curve

5. Summary

In this paper multiple regression models and logistic regression models have been investigated. Both categories of models are directly related and can be used for prediction, but they use different target variables. The primary objective of the study was to analyse how and to what extent prior information can influence the precision of regression and classification while using real data sets. First and foremost, the predictive analysis has been performed. This is because the outcomes of explanatory modelling cannot always be applied for predictive modelling (Provost and Fawcett, 2013).

To sum up, the predictive accuracy of models developed with non-informative and informative a priori distributions has been compared. The impact of prior information on the values of selected performance indicators developed for the models estimated with non-informative and informative a priori distributions has been shown. These results indicate that the accuracy of models estimated with informative a priori distributions is higher. Therefore, when additional out-of-sample knowledge is available, the appropriate selection of a priori distribution can improve the accuracy of regression and classification models.

REFERENCES

ALBERT, J. H., CHIB, S., (1993). Bayesian analysis of binary and polychotomos response data. Journal of the American Statistical Association, 88, 669–679.

BOLSTAD, W. M., (2007). Introduction to Bayesian statistics, USA: Wiley & Sons.

CONGDON, P., (2006). Bayesian Statistical Modelling, 2nd ed., UK: John Wiley & Sons Inc.

DRAPER, N., SMITH, H., (1981). Applied Regression Analysis, 2nd ed., New York: John Wiley & Sons.

FINNEY, D. J., (1972). Probit Analysis, London: Cambridge University Press.

GEWEKE, J., (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Bernardo J., Berger J., Dawiv A., Smith A. Bayesian Statistics, 4, 169–193.

GELMAN, A., CARLIN, J. B., STERN, H. S., RUBIN, D. B., (2000). Bayesian data analysis, London: Chapman & Hall/CRC.

GILL, J., (2008). Bayesian Methods, A Social and Behavioral Science Approach, USA: Chapman&Hall/CRC.

GOŁATA, E., (2004). Indirect Estimation of unemployment for the local labour market, Poznan: Publisher Academy of Economics in Poznan (in Polish).

GRZENDA, W., (2013). The significance of prior information in Bayesian parametric survival models. Acta Universitatis Lodziensis, Folia Oeconomica, 285, 31–39.

HOSMER, D. W., LEMESHOW, S., (2000). Applied Logistic Regression, New York: Wiley.

JAPKOWICZ, N., SHAH, M., (2011). Evaluating Learning Algorithms. A Classification Perspective, New York: Cambridge University Press.

KOOP, G., (2003). Bayesian Econometrics, Chichester, UK: Wiley.

LANCASTER, T., (2004). An Introduction to Modern Bayesian Econometrics, Oxford, UK: Blackwell Publishing.

PROVOST, F., FAWCETT, T., (2013). Data Science for Business: What You Need to Know About Data Mining and Data-analytic Thinking, USA: O'Reilly Media, Inc.

TUFFÉRY, S., (2011). Data Mining and Statistics for Decision Making, Chichester, UK: Wiley.

VEHTARI, A., OJANEN, J., (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. Statistics Surveys, 6, 142–228.