Erasmus University Rotterdam

Master Thesis in Applied Econometrics

Bayesian Variable Selection and Model Averaging for modelling the Probability of Default of mortgage portfolios

Author:
Luuk van Selm, 430382

Supervisor:
Prof. Martina Zaharieva (EUR)
Drs. Bojidar Ignatov (Deloitte)

July 16, 2019

Abstract

The subjective choice of which variables should be included in a Probability of Default (PD) model of a mortgage portfolio can significantly influence the outcome of the prediction. Bayesian Variable Selection (BVS) can be used to objectively estimate which variables should be included in the model, by assigning posterior probabilities to different variable combinations. Bayesian Model Averaging (BMA) can be used to average between these different combinations with the aim of decreasing the model specific errors. This paper investigates the improvements in prediction performance of PD models that can be achieved by implementing BVS and BMA. It is shown that BVS outperforms the benchmark variable selection criteria. It follows that implementing BMA results in a accuracy loss compared to specific BVS models, but results in more stability, robustness and overall more accurate predictions for selected model combinations.

Contents

1 Introduction

Previously issued financial instruments like mortgage loans are sold in secondary financial markets. Investors like the American Federal Home Loan Mortgage Corporation, also known as Freddie Mac, buy these loans with the goal of selling them as mortgage-backed securities. One of the most important factors of the valuation of these loans is estimating the Probability of Default (PD). Accurate models that predict the PD before a mortgage loan is bought, can be used as an acceptance rule or important valuation factor. Kotz (2009) states that the inability to correctly model these mortgage loans may have disastrous consequences, as was the case with the financial crisis of 2008. A widely used technique to model the PD is the logistic regression model, which calculates the relation between the PD of a mortgage loan and selected explanatory variables. Examples of these variables are a pre-calculated Credit Score, the Loan-To-Value Ratio or the Interest Rate of the loan. However, as it is not straightforward which variables should be included in the regression, the subjective choice of which to include can substantively influence the predictions of the model. Bayesian Variable Selection (BVS) can be used to estimate the optimal subset of included explanatory variables. The method uses Bayesian inference to calculate the posterior probability, which indicates the relative likelihood that the data is generated by the specified model, and can be used as a selection criteria. This makes it possible to objectively investigate which subset of explanatory variables should be used in a regression model. This paper starts with investigating the posterior probability of different subsets of explanatory variables, with the goal of estimating the optimal subset to be used in a PD model.

BVS assign's weights to different subsets of explanatory variables. In practice, this might result in different subsets that are comparably likely to be the optimal, but have significantly different parameter estimates. Choosing only one of these models incorporates model specific errors, that can be diversified by averaging between different potential models. This paper analyses Bayesian Model Averaging (BMA), which can be seen as an extension of BVS, and combines selected models weighted by their posterior probability as calculated with BVS. That way, different models containing different subsets of explanatory variables are combined with the goal of improving the prediction performance of the PD model. Consequently, the predicted PD becomes a weighted average of multiple models.

This paper analyses the corresponding changes in prediction performance obtained by the described Bayesian methods in comparison with traditional variable selection and model averaging methods, by answering the research question:
Does the use of Bayesian Variable Selection and Bayesian Model Averaging result in an increase in prediction performance when modelling the Probability of Default of mortgage portfolios?

The remainder of this section introduces the structure of the model, Bayesian inference, BVS, BMA, the prediction performance, a short summary of the results and this papers' contribution. The existing literature relevant for these topics is evaluated and the corresponding methods and criteria that best fit this research are selected.

Different modelling methods for the PD are intensively described in the existing literature. Beaver (1966) writes about models that predict defaults using financial ratios. In response, Alt-

man (1968) uses the dimension reduction technique called Multiple Discriminant Analysis to effectively model defaults, which is considered to be the most common technique for modelling PD in the years thereafter. Ohlson (1980) lies the foundation for the use of a logistic regression for these models. This method uses a logistic function to describe the relation between the explanatory variables and the PD. As Hosmer Jr et al. (2013) state, it has the advantages of being flexible, simplistic and gives a meaningful interpretation to the obtained results. Furthermore, it is the most commonly used technique in the contemporary literature. This paper uses the logistic regression model to describe the PD. As Menard (2002) states, the Ordinary Least Squares estimation assumptions of homoscedasticity, linearity and normality are likely to be violated for this type of regression, resulting in inefficient estimates. Therefore, after completing BVS, this paper estimates the selected models through Iteratively Re-weighted Least Squares (IRLS), as described by Green (1984), that uses Maximum Likelihood as estimation method for the parameters. Hence, after the use of the Bayesian methodology, a frequentist approach is used to estimate the models. This method is selected as Bayesian methodology, including Bayesian estimation, is not commonly used in the field of credit risk. Hence it is investigated whether the variable selection techniques can improve the existing PD models, instead of whether they can be used for models that are not used in practise. As IRLS is commonly used to estimate logistic regression model in the field of credit risk, the obtained results are generally applicable. In short, it is investigated if the use of BVS and BMA can improve models based on the current estimation methods used in the field of credit risk.

The foundation of both BVS and BMA lies in Bayesian inference, that applies probabilities to statistical problems. Prior believes about the distribution of vector $\boldsymbol{\theta}$, containing the parameters of interest, are denoted by $\pi(\boldsymbol{\theta})$. They are updated by likelihood of the data D, as a function of these parameters, denoted by $\pi(D|\boldsymbol{\theta})$. The product is proportional to the posterior beliefs $\pi(\boldsymbol{\theta}|D)$ about the parameters given the data. This is written as

$$\pi(\boldsymbol{\theta}|D) \propto \pi(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{1}$$

When explaining these three different components of Bayesian inference, we start with the analysis of a single model. Furthermore, when moving on to BVS and BMA, the aim is to compare different models.

Bayesian inference starts with specifying prior believes about the parameters $\pi(\boldsymbol{\theta})$. They are called the prior distribution, a set of initial beliefs about the parameters $\boldsymbol{\theta}$ without observing the data. A broad selection of differently shaped priors is available in literature. Different industries, topics or regression types result in different prior beliefs about the parameters. Consequently, the use of a logistic regression model results in its own set of potential priors. Fragoso et al. (2018) gives an overview of the available options analysed in literature. As described above, after obtaining the BVS results, the logistic regression is estimated by using IRLS, which does not require priors. Carefully selected priors are however required to enhance the effectiveness of BVS. Most commonly used are the so called spike-and-slab priors. The basic idea of these priors is comparable with the classical Lasso and Ridge regressions (Owen, 2007). By shrinking the coefficients of the variables that are less relevant when describing the dependent variable, they highlight the difference in importance between the explanatory variables. George & McCulloch

(1993) describe a set of parameter priors that are based on this concept and adapted to be used with the efficient sampling method called Stochastic Search Variable Selection (SSVS). This prior significantly adds to the effectiveness of BVS and is required for an effective implementation of the SSVS technique, which is used in this paper. This paper uses a version of this prior, which is modified to be compatible with the use of a logistic regression and tuned to fit to the data set.

Next, we specify the likelihood function $\pi(D|\boldsymbol{\theta})$. In order to do so we have to model the probability of obtaining the observed data, given the specified parameters. This is done by specifying the model and the corresponding distribution of its parameters. As described above, the PD is described by using a logistic regression. As stated by Czepiel (2002), the likelihood function corresponding to this type of regression is based on the Binomial distribution.

Analytically calculating the posterior distribution $\pi(\boldsymbol{\theta}|D)$ can be difficult. The use of a non-conjugate prior, a prior that has a different distribution than the posterior, also significantly increases the computational power required to obtain the posterior distribution. Moreover, for some model specifications, like for example the logistic regression, it is not possible to analytically obtain marginal posterior results. Monte Carlo Markov Chain (MCMC) methods, like the Gibbs sampler and Metropolis-Hasting algorithms (Roberts & Smith, 1994), can be used to simulate from posterior distribution. The literature also introduces new, more robust sampling methods like Hamiltionian Monte Carlo (Hoffman & Gelman, 2014). These sampling methods can be used to sample from the distribution of the posterior.

The posterior distribution of a single model can be used for predicting. Posterior distributions of multiple models can be used to estimate posterior probabilities, that are used as variable selection criteria for BVS. The posterior probability describes the relative probability that one model best describes the data by evaluating the posterior probability. By analyzing every possible combination of potential explanatory variables as a separate model, the subsets are ordered on their posterior probability. This paper evaluates the relative effect of BVS by comparing prediction performance of the obtained top combinations with the traditional variable selection criteria Bayesian Information Criteria (BIC) and Aikake Information Criteria (AIC). For p explanatory variables, there are $k = 2^p$ different possible combinations. Because every different combination implies a new model, this results in k different models $M_l, ..., M_k$. In this research we investigate the potential inclusion of 9 different explanatory variable, resulting in $2^9 = 512$ different models. Estimating the posterior probability of all of these models is inefficient, since there is solely an interest in the models that have a large posterior probability. Therefore, researchers developed a wide range of BVS algorithms that deal with this inefficiency. The most commonly used methods are the Stochastic Search Variable Selection (SSVS) algorithm, first described by George & McCulloch (1993), and the Monte Carlo Markov Chain Reverse Jump (MCMCRJ) algorithm, as stated by Green (1995). SSVS uses the Gibbs sampler in combination with the earlier described spike-and-slab prior, in order to identify the models with higher posterior probabilities. MCMCRJ has a different approach. It first analyses the number of included explanatory variables, and secondly determines which ones to include specifically using the Metropolis Hastings sampler. The SSVS technique is used, which is in line with the use of the SSVS prior as stated earlier. The effectiveness of its application is comparable to that

of MCMCRJ (O'Hara et al., 2009), but the technique is more effective given the used prior.

BMA combines different models, weighted by their posterior probability, as described by Fragoso et al. (2018). This is the same criteria as used by BVS. Madigan & Raftery (1994) state that this model averaging method increases the prediction performance, in comparison to the use of a single model. There is no general consensus in existing literature about the number of models that should be included in the weighted combination. Madigan & Raftery (1994) describe two main rules of inclusion for models. The first is that models that predict the data 'far less well' than the optimal model should not be included. Secondly, a model should not be included if a subset of the models' parameters has a higher prediction performance. This paper investigates the effectiveness of the application of these rules separately and together. Additionally, this paper experiments with different combinations of selected models, with the aim of finding the combination that has the overall highest prediction performance for this research. This is done using the top 10 models with the highest posterior probability obtained by BVS. The selected different combinations are evaluated on their prediction performance, and compared to the results of the individual models obtained through BVS.

Literature shows a large range of different weights that can be used when combining models, often proportional to some calculated model selection criteria. In addition to the posterior probability, Zhang et al. (2006) state that the Aikake Information Criteria (AIC) or the Bayesian Information Criteria (BIC) can be used as weight. Another option is a simple average, where the weights of all h models are set equal at $\frac{1}{h}$. This paper will use these more traditional weights as a benchmark.

The prediction performance of the different models is measured in terms of the model's ability to discriminate between defaults and non-defaults, and the prediction accuracy. The discriminatory performance is quantified using the Accuracy Ratio (AR), and the Area Under the Curve (AUC) as described by Lingo & Winkler (2008). Abdi (2007) describes the Binomial test, that quantifies the model's accuracy. These are some of the industry standard backtesting measures.

The results obtained in this paper show that BVS does outperform the traditional criteria and algorithms that are used as a benchmark, and hence results in an increase in prediction performance. Furthermore, carefully selected BMA combinations on average have a better prediction performance than BVS, and also shows less variation in the model specific errors. However, BMA does not significantly outperform the traditional weighting benchmark. Hence averaging the through BVS obtained models, with for example BMA, results in combinations with improved prediction performance.

Literature gives a broad description of different applications of BVS and BMA for logistic regressions. However, its application to credit risk is limited, and the focus on the PD of single family household mortgage loans is a new concept. Additionally there are few examples of the application of SSVS for a logistic regression, none of them focusing on the topic of credit risk. This paper contributes to existing literature by quantifying the changes in prediction prediction performance as a consequence of BVS and BMA. Furthermore, the comparison with more traditional variable selection and model averaging methods such as AIC and BIC is a new concept in the field of credit risk. The same goes for the analysis of the relation of the

PD with the selected optimal subset of explanatory variables as estimated with BVS. Lastly, this research will investigate the effectiveness of the few available guidelines for the number of included models in BMA, and investigates potential new and more robust solutions.

The remainder of this paper has the following structure. Section 2 describes the methodology behind the research, stating and elaborating on the formulas describing the PD, Bayesian inference, BVS, BMA, the prediction performance criteria, and describing the used programming tools. Section 3 describes the data set, and explains how the data is used in this research. Section 4 describes the obtained results of BVS and BMA, and compares them to each other. Lastly, Section 5 states the conclusions and provides recommendations for the direction of further research.

2 Methodology

In this section, the methodology is presented. Section 2.1 starts with a description of the logistic regression model that is used to model the PD. Section 2.2 and 2.3 give an overview of Bayesian inference, and elaborate on the criteria used to compare different models, e.g., Bayes factors, posterior odds and the posterior probability. Additionally, the link to the posterior probability is shown. Section 2.4 elaborates on BVS and different traditional variable selection techniques. It describes SSVS and the selected prior distribution, the likelihood, and the posterior distribution. Section 2.5 describes BMA and different traditional weights that can be used for model averaging. Section 2.6 defines prediction performance and the corresponding performance selection criteria that are used in this paper. Section 2.7 gives an overview of the described algorithms and criteria and explains how the different results are compared. Section 2.8 describes the programming tools and packages used in this paper. Finally, Section 2.9 describes the assumptions made throughout this paper. Figure 1 gives a graphical overview of the methodology section.

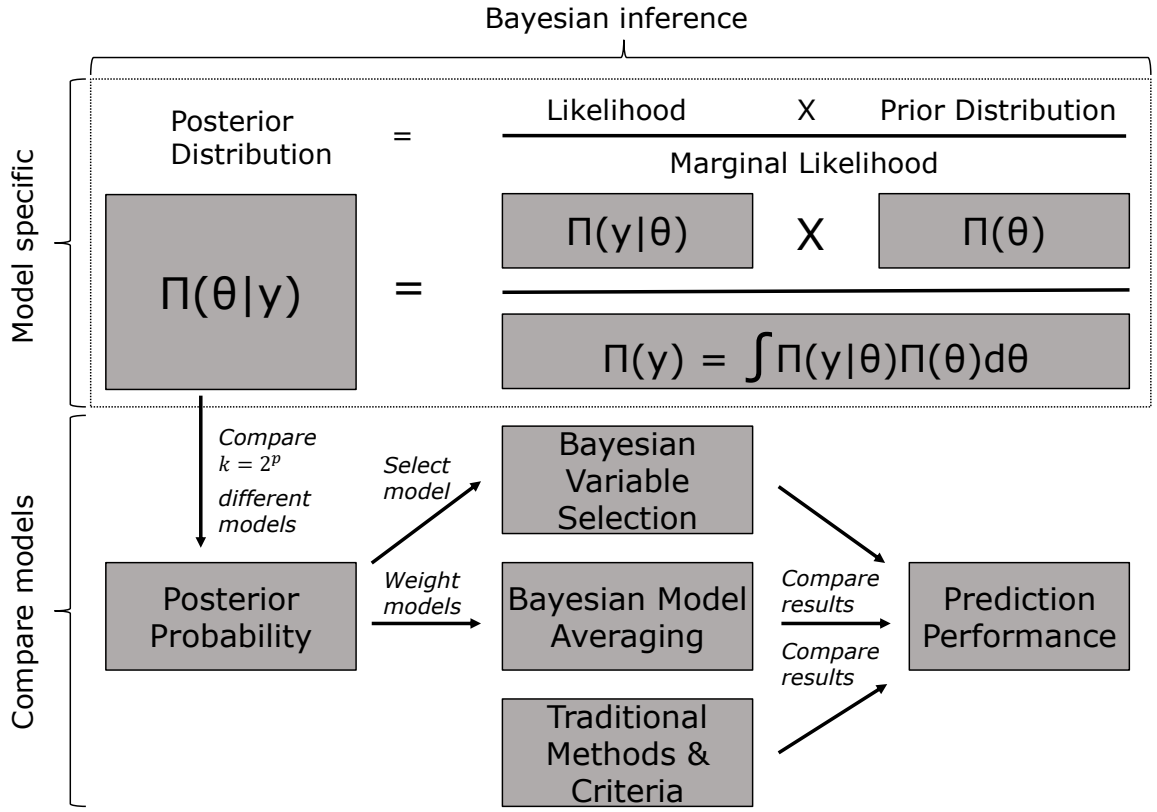


Figure 1: An overview of the build-up of the Methodology section. Starting with model specific Bayesian inference, thereafter explaining the link to comparing models with the Posterior Probability, selecting them with Bayesian Variable Selection and combining them with Bayesian Model Averaging. Finally the results are compared on prediction performance, where traditional methods using traditional criteria are used as a benchmark.

In Figure 1, $\boldsymbol{\theta} = (\theta_1, ..., \theta_g)$ indicates a vector including the g parameters of the model, and $\boldsymbol{y}$ is a $(n \times 1)$ vector containing the binary dependent variable indicating a default as 1, and a non-default as 0.

2.1 The Logistic PD Model

This paper uses a logistic regression to model the PD. Based on the description in Hosmer Jr et al. (2013), we define the PD as the mean of $\boldsymbol{y}$ conditional on $\boldsymbol{X}$. That is $\pi(x) = E(\boldsymbol{y}|\boldsymbol{X})$, where $\boldsymbol{X}$ is a $(n \times p + 1)$ matrix containing a constant and p explanatory variables. The defaults can be expressed as $\boldsymbol{y} = \pi(x) + \varepsilon$, where $\boldsymbol{y} \sim Bi(n, \pi(x))$. $\pi(x)$ indicates the PD conditional on the explanatory variables, and can be expressed as

$$\pi(x) = E(\boldsymbol{y}|\boldsymbol{X}) = F(\boldsymbol{\beta'x}) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}. \tag{2}$$

Here, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$, where β_0 indicates the constant, and β_i represents the regression coefficient for explanatory variable x_i. Using a logit transformation we obtain

$$g(x) = ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p, \tag{3}$$

where the parameters of $g(x)$ are linear, and the function has a range of $(-\infty, \infty)$. The likelihood function corresponding to the Binomial distribution of the logistic regression is of the form

$$\pi(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x)^{y_i}(1-\pi(x))^{(1-y_i)}, \tag{4}$$

where $y_i = 1$ indicates a default of observation i, and $\pi(x_i)$ is the probability of this default for observation i. By substituting Equation 2 into Equation 4, we obtain the likelihood function

$$\pi(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\left(\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}} \right)^{y_i} \left(1 - \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}} \right)^{(1-y_i)} \right]. \tag{5}$$

After selecting the optimal set of explanatory variables, the parameters are estimated using IRLS. As described in Green (1984), this method iteratively uses the standard Newton-Raphson algorithm to obtain the maximum likelihood estimates for Equation 5.

2.2 Bayesian Inference

The Bayesian inference model used in this research is described in Raftery (1995). The prior beliefs about the unknown parameter vector $\boldsymbol{\beta}$ of a single model, are described by the probability density function $\pi(\boldsymbol{\beta})$. The likelihood function is described as $\pi(\boldsymbol{y}|\boldsymbol{\beta})$. The product of the likelihood function and the prior beliefs is proportional to the posterior distribution of the parameters conditional on the data, as

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})}{\pi(\boldsymbol{y})} = \frac{\pi(\boldsymbol{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})}{\int \pi(\boldsymbol{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})d\boldsymbol{\beta}} \propto \pi(\boldsymbol{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})). \tag{6}$$

The equality in the middle is obtained by the law of total probability, which states

$$\pi(\boldsymbol{y}) = \int \pi(\boldsymbol{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})d\boldsymbol{\beta}. \tag{7}$$

This is called the marginal likelihood and is the cornerstone of Bayesian model comparison and hence variable selection. Proper density functions, that integrate to 1, are required for model selection in order to make unbiased comparisons between models. By obtaining the marginal likelihood, the product of the likelihood function and a proper prior can be transformed into a proper density function.

2.3 Bayes Factors

Bayes factors are used to quantify the probability that a model is correct, in comparison to one or more different models. For model l, we define the marginal likelihood function of the model as

$$\pi_l(\boldsymbol{y}) = \int \pi_l(\boldsymbol{y}|\boldsymbol{\beta})\pi_l(\boldsymbol{\beta})d\boldsymbol{\beta}. \tag{8}$$

Now the Bayes factor of model l to model j, can be obtained by

$$BF_{lj} = \frac{\pi_l(\boldsymbol{y})}{\pi_j(\boldsymbol{y})}, \tag{9}$$

where $BF_{lj} \geq 1$ states that M_l is equally or more likely than M_j given that both models are equally likely a priori. Casella & George (1992) describe that, when one model is a priori more likely than another model, the ratio of the prior likeliness can be used to calculate the posterior odds of the data supporting M_l over M_j as $Posterior\ Odds = Prior\ Odds \times Bayes\ Factor$, or

$$\frac{Pr(M_l|\boldsymbol{y})}{Pr(M_j|\boldsymbol{y})} = \frac{Pr(M_l)}{Pr(M_j)} \times BF_{lj}, \tag{10}$$

where $\Pr(M_l)$ states the prior probability that M_l is the correct model. Note that the prior odds depend on the prior distribution of gamma as specified in Equation 20. A lower prior probability of including a variable results in a preference for models with less explanatory variables.

By calculating the Bayes factors of all the potential models against a selected baseline model M_1, the Bayes factor between any two models can be calculated. For M_l, the posterior odds against M_j are calculated using both models Bayes factors with the baseline model. This is done with the formula

$$BF_{ij} = \frac{\pi_l(\boldsymbol{y})}{\pi_j(\boldsymbol{y})} = \frac{\pi_l(\boldsymbol{y})}{\pi_1(\boldsymbol{y})} \times \frac{\pi_1(\boldsymbol{y})}{\pi_j(\boldsymbol{y})} = BF_{i1} \times \frac{1}{BF_{j1}} = \frac{BF_{l1}}{BF_{j1}}. \tag{11}$$

This can be done in a similar manner for the prior and posterior odds.

By comparing the posterior odds of a model against all other potential models, the posterior probability is obtained. As stated by Fragoso et al. (2018), for k different models, and $l \in 1, ..., k$, we estimate

$$Pr(M_l|\boldsymbol{y}) = \frac{\pi_l(\boldsymbol{y})Pr(M_l)}{\sum_{j=1}^{k} \pi_j(\boldsymbol{y})Pr(M_j)}, \tag{12}$$

with the assumption that all models are equally likely a priori and by substituting Equation 9 and Equation 11 into Equation 12, we obtain the formula for the posterior probability of model l in terms of the Bayes factors, that is

$$Pr(M_l|\boldsymbol{y}) = \frac{BF_{l1}}{\sum_{j=1}^{k} BF_{j1}}. \tag{13}$$

This posterior probability is the criteria used in both BVS and BMA. This relation works two ways. When we estimate the posterior probability of a model through simulation, Equation 13 can be used to calculate the Bayes factors. This can be used to quantify the relative probability between two or multiple models.

2.4 Variable Selection

The goal of variable selection is to investigate whether different explanatory variables should be included in a model. On one hand, adding a variable can increase the accuracy of the model. On the other hand, it will also increase the complexity of a model and the possibility of over-

fitting on some aspects of the data set. Therefore, the model's optimal ratio between being accurate and being parsimonious should be evaluated. There are multiple methods available that investigate this ratio using different algorithms and selection criteria. This paper will analyse the BVS method in its ability to calculate the optimal subset of explanatory variables, and uses traditional variable selection methods as a benchmark.

2.4.1 Bayesian Variable Selection

BVS uses the posterior probability, calculated in Equation 12, as selection criteria. This paper uses the SSVS algorithm to select the models with the largest posterior probability for BVS. The following sections will elaborate on the SSVS algorithm, including the Gibbs sampler, the selected priors, likelihood function and corresponding posterior distribution. Lastly, traditional benchmark methods and criteria are described and selected.

2.4.1.1 Stochastic Search Variable Selection

Calculating the posterior distribution of all potential models is inefficient, since we are only interested in the models with relatively large posterior probabilities. As mentioned in Section 1, this paper uses the SSVS algorithm as described by George & McCulloch (1993), to only select the models with a large posterior probability. Furthermore, the algorithm uses the Gibbs sampler to simulate from this posterior probability.

We introduce the latent binary variable $\gamma_i \in \boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)$, which has value 1 for including x_i, and value 0 for excluding x_i in the model. Conditionally γ_i, β_i is distributed as the normal mixture model

$$\beta_i|\gamma_i \sim (1-\gamma_i)N(0, \tau_i^2) + \gamma_i N(0, c_i^2\tau_i^2). \tag{14}$$

This distribution is based on the concept of spike-and-slab as stated in Section 1. Here, τ_i is small, so if $|\beta_i| \leq 3\tau_i$, it is assumed that $\beta_i = 0$ and the explanatory variable should not be included in the optimal subset. c_i is larger, stating that if $|\beta_i| \geq 3c_i\tau_i$, the corresponding explanatory variable should be included in the optimal subset. Hence the sizes of c_i and τ_i are data specific and different options are evaluated in Section 4. Consequently, the probability that β_i has an estimate that differs from 0 and hence x_i should be included in the optimal subset of variables and is denoted as

$$P(\gamma_i = 1) = 1 - P(\gamma_i = 0) = p_i. \tag{15}$$

The vector $\boldsymbol{\gamma}$ contains p random variables that are either 0 or 1, indicating whether the corresponding variables are included. Hence, the different vectors of $\boldsymbol{\gamma}$ indicate different combinations of explanatory variables and consequently different models $M_1, ..., M_k$.

2.4.1.2 The Gibbs Sampler

First off, the separate distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, conditional on the data and the other parameter, are obtained. Secondly, a Markov chain of the parameters is constructed from which the limit of the distribution converges towards the marginal posterior distribution of the parameters. This

'Gibbs sequence' has the form

$$\boldsymbol{\beta}^0, \boldsymbol{\gamma}^0, \boldsymbol{\beta}^1, \boldsymbol{\gamma}^1 ..., \boldsymbol{\beta}^j, \boldsymbol{\gamma}^j, .., \tag{16}$$

where $\boldsymbol{\gamma}^j$ is the jth sample of $\boldsymbol{\gamma}$. An iterative procedure is used to sample from these distributions. The used method is called the Gibbs sampler, and sequentially updates the parameters by sampling $j = 1, ..., m$ iterations from the conditional distributions of the parameters. Raftery & Lewis (1991) state that 1000 to 5000 iterations are required, therefore we set $m = 5000$ for this paper. Additionally, we make use of a burn in period of 500 observations. For $\boldsymbol{\beta}^j$ and $\gamma_i^j \in \boldsymbol{\gamma}^j = (\gamma_1^j, ..., \beta_p^j)$, we sample

$$\begin{aligned} \boldsymbol{\beta}^j \sim f(\boldsymbol{\beta}^j | \boldsymbol{\beta}^{j-1}, \boldsymbol{y}, \boldsymbol{\gamma}^{j-1}) \propto \pi(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) \times \pi(\boldsymbol{\beta}^{j-1}|\boldsymbol{\gamma}^{j-1}), \\ and\ \gamma_i^j \sim f(\gamma_i^j | \boldsymbol{\beta}^j, \boldsymbol{\gamma}_{(i)}^j) \propto Bernoulli\left(\frac{a^j}{1+a^j}\right), \\ where\ a^j = \frac{\pi(\boldsymbol{\beta}|\gamma_i = 1, \boldsymbol{\gamma}_{(i)}^j)}{\pi(\boldsymbol{\beta}|\gamma_i = 0, \boldsymbol{\gamma}_{(i)}^j)} \times \frac{\pi(\gamma_i = 1, \boldsymbol{\gamma}_{(i)}^j)}{\pi(\gamma_i = 0, \boldsymbol{\gamma}_{(i)}^j)}, \end{aligned} \tag{17}$$

where $\boldsymbol{\gamma}_{(i)}^j = (\gamma_1^j, ..., \gamma_{i-1}^j, \gamma_{i+1}^{j-1}, ..., \gamma_p^{j-1})$. Note that the distribution of γ_i^j is independent of $\boldsymbol{y}$. The generated sequence $\boldsymbol{\gamma^1}, ..., \boldsymbol{\gamma^m}$ rapidly converges to the marginal posterior distribution $\boldsymbol{\gamma} \sim f(\boldsymbol{\gamma}|\boldsymbol{y})$. In this sequence, the $\boldsymbol{\gamma}$'s that imply a model with a higher posterior probability will come up more often. By simply counting the appearances of different $\boldsymbol{\gamma}$'s, the models can be ranked on posterior probability. We say that $\#\boldsymbol{\gamma}_{M_l}$ equals the amount of times that we sample the vector $\boldsymbol{\gamma}_{M_l}$, that is the vector indicating the inclusion set of explanatory variables that result in the model M_l. Consequently, the posterior probability of M_j can be estimated as

$$Pr(M_l|\boldsymbol{y}) = \frac{\#\boldsymbol{\gamma}_{M_l}}{m}. \tag{18}$$

The priors and likelihood functions from which the samples are taken are evaluated in the next section.

2.4.1.3 The Prior Distribution

Prior distribution gives the possibility to incorporate a priori known information about the parameters into the model. For the logistic regression, we need to specify our prior believes about the parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. That is $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma})$, and $\pi(\boldsymbol{\gamma})$. As stated by George & McCulloch (1993), the prior of $\boldsymbol{\beta}$ should be be used to shape the distribution in a form that is compatible with SSVS. Furthermore, this prior includes scaling variables that can be used to adapt the models to the specific research and data requirements. For $\boldsymbol{\beta}$, a multivariate normally distributed prior is selected, which is defined as

$$\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) \sim N_p(0, \boldsymbol{D_\gamma R D_\gamma}), \tag{19}$$

where $\boldsymbol{R}$ is the correlation matrix of the prior distribution and $\boldsymbol{D_\gamma} \equiv diag[a_1\tau_1, ..., a_p\tau_p]$ where $a_i = 1$ for $\gamma_i = 0$, and $a_i = c_i$ for $\gamma_i = 1$. With the use of this prior, the mixture model as stated in Equation 14 is obtained. Here, τ_i and c_i are fitted to the data as explained in Section

2.4.1.1. We set $\boldsymbol{R} = I$, which implies that the $\boldsymbol{\beta}$ conditional on $\boldsymbol{\gamma}$ is independent.

For $\boldsymbol{\gamma}$, the choice of prior should include any information that is known a priori about which explanatory variables should be included in the optimal subset. Two options, stated by George & McCulloch (1993), are investigated. These are

$$\begin{aligned} \pi(\boldsymbol{\gamma}) &= \prod_{i=1}^{n} p_i^{\gamma_i}(1-p_i)^{(1-\gamma_i)}, \\ &\quad and\ \pi(\boldsymbol{\gamma}) = 2^{-p}. \end{aligned} \tag{20}$$

The first prior implies independent γ_i's. We will investigate the consequences of different values for p_i in Section 4. The second prior is a special case of the first one, stating that the probability of inclusion for every explanatory variable is equal to $p = \frac{1}{2}$.

2.4.1.4 The Likelihood Function

The likelihood function is similar to Equation 5. Additionally, $\boldsymbol{y}$ now also depends on $\boldsymbol{\gamma}$ through $\boldsymbol{\beta}$. This can be stated as

$$\pi(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\gamma}) = \prod_{i=1}^{n}\left[\left(\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}\right)^{y_i}\left(1-\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}\right)^{(1-y_i)}\right]. \tag{21}$$

2.4.1.5 The Posterior Distribution

As stated in Equation 6, the posterior distribution is proportional to the product of the likelihood and the priors. In this hierarchical Bayesian model, the posterior distribution can be written as

$$\pi(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\gamma}) \times \pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) \times \pi(\boldsymbol{\gamma}). \tag{22}$$

The three functions on the right are stated in Equation 21, Equation 19, and 20. Hence, by substituting these equations into Equation 22, we obtain

$$\begin{aligned} \pi(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{y}) \propto \prod_{i=1}^{n}\Bigg[&\left(\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}\right)^{y_i}\left(1-\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}\right)^{(1-y_i)} \\ &\times\frac{1}{(2\pi)^{\frac{2}{p}}(\boldsymbol{D_\gamma R D_\gamma})^{\frac{1}{2}}}\exp\Big\{-\frac{1}{2}\boldsymbol{\beta' D_\gamma R D_\gamma \beta}\Big\} \times p_i^{\gamma_i}(1-p_i)^{(1-\gamma_i)}\Bigg], and \\ \pi(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{y}) \propto \prod_{i=1}^{n}\Bigg[&\left(\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}\right)^{y_i}\left(1-\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}\right)^{(1-y_i)} \\ &\times\frac{1}{(2\pi)^{\frac{2}{p}}(\boldsymbol{D_\gamma R D_\gamma})^{\frac{1}{2}}}\exp\Big\{-\frac{1}{2}\boldsymbol{\beta' D_\gamma R D_\gamma \beta}\Big\} \times 2^{-p}\Bigg]. \end{aligned} \tag{23}$$

Consequently, the posterior distributions are made proper by dividing them by their marginal likelihood as calculated with Equation 8. In this formula, model i relates to the explanatory variable combination as stated by the corresponding $\boldsymbol{\gamma}$. As stated earlier, the Gibbs sampler of the SSVS algorithm is used to sample from this distribution and obtain the part that we are

interested in, that is the posterior marginal distribution described by $\boldsymbol{\gamma} \sim f(\boldsymbol{\gamma}|\boldsymbol{y})$. This is used to estimate the posterior probability as stated in Equation 18.

2.4.2 Traditional Variable Selection

Three algorithms that are commonly used for variable selection are Forward Selection (FS), Backwards Eliminations (BE), and Stepwise Selection (SS). FS starts with a model without explanatory variables. Given a selection criteria, it investigates which variable should be added to the model. This continues until the addition of another variable significantly decreases the selected criteria. Now the model, consisting of the subset of variables with the highest selection criteria, is selected. BE follows the same principle, but starts with all available variables and eliminates them one by one from the regression with the goal of optimizing the selection criteria in every step. SS is a hybrid of the two other algorithms, considering both eliminating and including potential variables in every step given selected criteria.

Raftery (1999) states that for a linear regression with specifically specified priors, the BIC approximates the Bayes factors. As the regression used in this paper is logistic, the results obtained when using the BIC criteria can be used as a benchmark. Additionally, AIC is selected. This paper uses these criteria in combination with the FS, BE, and SS algorithms as a benchmark for BVS.

2.5 Model Averaging

Model averaging is used to create a weighted average that diversifies the model specific errors. By doing so, this paper aims to create a combination of models that has increased prediction performance. A weighted combination of h selected models is used to create an averaged estimate of the conditional PD as calculated in Equation 2. This is described by

$$\bar{\pi(x)} = w_1\pi(x)_1 + w_2\pi(x)_2 + ... + w_h\pi(x)_h, \tag{24}$$

where w_1 indicates the weight, and $\pi(x)_1$ the conditional PD, corresponding to Model 1. The sum of the weights equals 1. The following sections will elaborate on BMA. Furthermore, the effects of traditional model averaging weights are analysed as a benchmark.

2.5.1 Bayesian Model Averaging

BMA creates a weighted combination of h models with a high posterior probability. For the weights, the posterior probability of model i relative to the other included models, is calculated as

$$w_{BMA_i} = \frac{BF_{l1}}{\sum_{j=1}^{h} BF_{j1}}. \tag{25}$$

This equation differs from Equation 13, that is used as criteria for BVS, in that only the h models that are included in the combined average are taken into account, in contrast to including all k models. Consequently, the weights are not equal to the posterior probabilities as described by Equation 13, but proportional to them, constrained by the rule that their sum is 1. In short, the same criteria are used for both BVS and BMA, but for BMA its value is proportionally corrected

for the number of models combined. As stated in Section 1, there is no consensus about the number of models that should be combined. This paper will analyse different combinations of models and evaluate them on their prediction performance. Possibilities for a 'rule of thumb' for the number of combined models are investigated.

2.5.2 Traditional Model Averaging Weights

In literature, there is a large amount of potential weights available. Most of them are based on variable selection criteria. This paper will analyse three different options: Equal weights, AIC and BIC. Equal weights indicates that all h models are assigned the same weight, that is

$$w_{eq} = \frac{1}{h}. \tag{26}$$

The AIC and BIC weights are based on the inverse of their relative criteria size, as described by Posada & Buckley (2004). For model i, this corresponds to

$$\begin{aligned} w_{AIC_i} &= \frac{\frac{1}{AIC_i}}{\sum_{i=1}^{h} \frac{1}{AIC_j}}, \quad where\ AIC_i = 2g_i - 2ln(\hat{L}_i), \\ and\ w_{BIC_i} &= \frac{\frac{1}{BIC_i}}{\sum_{i=1}^{h} \frac{1}{BIC_j}}, \quad where\ BIC_i = ln(n)g_i - 2ln(\hat{L}_i), \end{aligned} \tag{27}$$

where g_i indicates the number of parameters estimated by model i. For the logistic regression, this equals the p regression coefficients of the model specific explanatory variables, a constant and the variance of the error term. L_i indicates the maximum value of model i's likelihood function and n equals the number of observation. Smaller values of the criteria result in larger weights for the corresponding model, hence the weights are proportional to the model's likelihood, penalized by the amount of parameters. For $n \geq 4$, the penalty on an additional parameter is larger for BIC, resulting in lower weights for larger models.

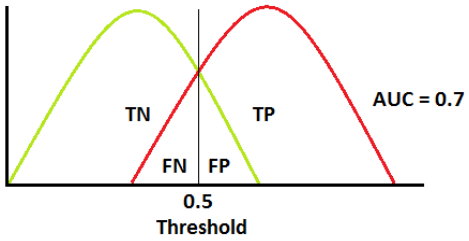
2.6 Evaluating Prediction Performance

By splitting the data set in two, a train and a test set are created. The test set is used to validate the out-of-sample prediction performance of the models. This is done by analyzing a model's ability to discriminate between defaults and non-defaults, and its prediction accuracy. The posterior probability of a model is also a criteria that can be used to select models on their prediction performance. Hence models with a higher posterior probability are expected to have a higher prediction performance. The used prediction performance criteria are a modified version of the Accuracy Ratio (AR), the Area Under the Curve (AUC), and the Binomial test. The first two tests indicate a model's discrimination performance, whereas the last test indicates the model's accuracy.

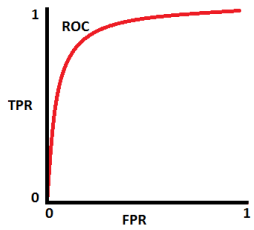
After predicting the probability of default of the different portfolios, a threshold is set that indicates for which probabilities a portfolio is predicted to default. This results in a binary variable indicating a predicted default or non-default. When the amount of defaults are approximately equal to the non-defaults, an optimal threshold can be calculated, for which

both increasing and decreasing the threshold would result in lower discriminatory power of the model. In this paper, the amount of defaults is relatively low, consequently the optimal threshold would qualify all observations as a non-default. A loss function, putting more weight on a falsely predicted default (FN), is a possible solution to this problem. However, as there is no general economically interpretable solution for the difference in weight, the subjective choice of a loss function can influence the general application of the results. Hence we make use of another threshold, indicating the minimum percentage of correctly predicted defaults, also known as the True Positive Rate (TPR). The amount of FP, given the minimum percentage of TPR is compared between the different models. This is done for thresholds of 60%, 90% and the average in the range of 50% to 100%. Additionally for some models, the amount of wrongly predicted non-defaults is plotted against this threshold ranging from 0 to 1. This measurement can be considered a modified version of the AR, as the AR simply states the ratios of correct predicted defaults, false predicted defaults, correct predicted non-defaults and false predicted non-defaults.

The AUC tests also models the ability to discriminate between defaults and non-defaults. By plotting the defaults and non-defaults separately on their predicted PD, Figure 2a is obtained. Additionally to the above described abbreviations, TP indicates a correctly predicted default, and TN corresponds to a correctly predicted non-default. The threshold for the PD indicates that portfolios with a higher PD are predicted to default. The ratio between TP and FP is plotted for all possible thresholds. The curve in this figure is known as the Receiver Operator Curve (ROC) and displayed in Figure 2b. The area underneath this curve ranges from 0 to 1, where a larger value indicates that the model is better able to discriminate between defaults and non-defaults. Specifically, 1 states a perfect ability to separate, 0.5 states not being able to separate and 0 states that the model reciprocates the classes.



(a) Two distributions actual defaults and non-defaults, plotted against their predicted PD. The threshold indicates the criteria that is used to predict whether a portfolio defaults or not. In this case, if the probability of default exceeds 0.5 a default is predicted. TP indicates the correctly predicted defaults and FP the falsely predicted defaults. Similar, TN corresponds to correctly predicted non-defaults and FP to falsely predicted non-defaults.



(b) The ROC curve. Displays the True Positive Ratio (TPR) on the Y-axis, indicating the correctly predicted defaults. On the X-axis, the False Positive Ratio (FPR) is displayed, indicating the wrongly predicted defaults. The range of both axes indicates the threshold, ranging from 0 to 1. The area underneath the curve indicates the models discriminatory power.

The Binomial test approaches the binomial distribution of the defaulting observations by a normal distribution, and investigates whether the amount of predicted defaults significantly

differs from the actual defaults. Using a Z-test, we test

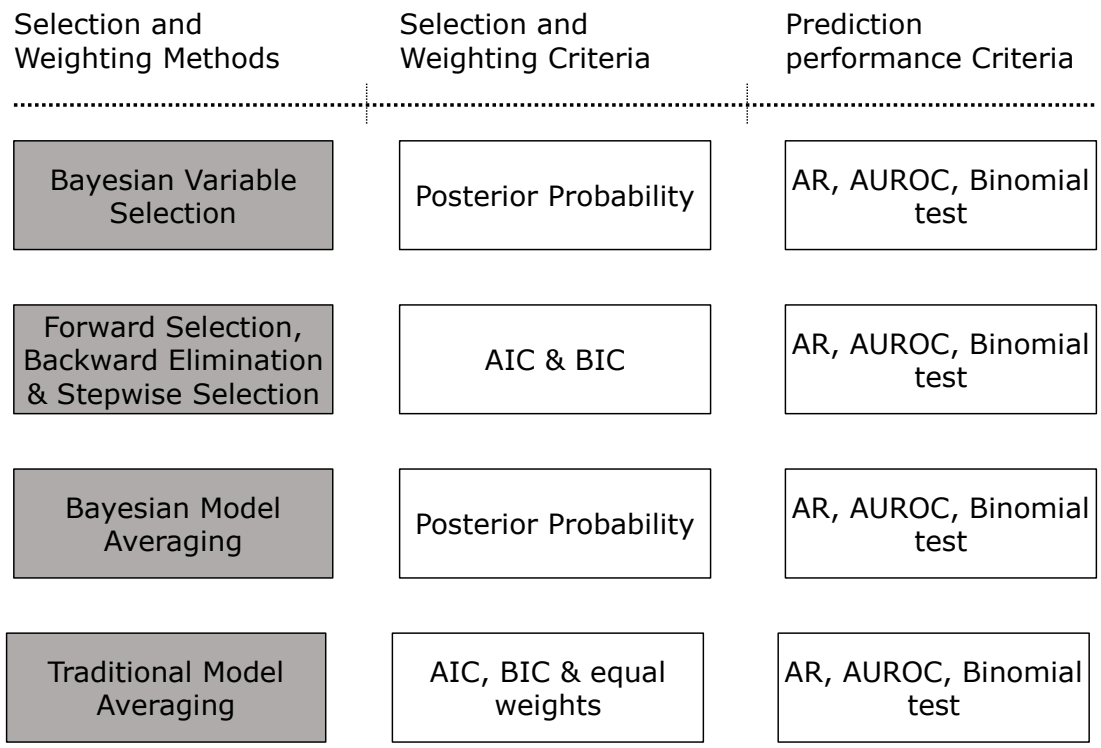
$$Z = \frac{n\pi(x) - \sum_{i=1}^{n} y_i}{\sqrt{n\pi(x)(1-\pi(x))}}. \tag{28}$$

It follows that, for a 95% confidence interval, $Z \leq 1.96$ indicates that the predicted number of defaults does not significantly differ from the actual number of defaults. As a consequence of setting the TPR at a high percentage, all models will have Z-values that lie outside the 95% confidence interval. Nevertheless the estimated Z-value can be considered a criteria indicating the accuracy of the amount of predicted defaults normalized by the models variance, and hence compared between the different models.

2.7 Comparing Models Set-up

Figure 8 gives an overview of the methods and criteria analysed and compared in this paper. For BVS, different options for c_i and $f(\boldsymbol{\gamma})$ and their corresponding selected models are first investigated. Furthermore, the top 5 models with the highest posterior probability is analysed in terms of prediction performance and their traditional criteria values. Next the results are compared to the models created with the FS, BE, and SS algorithms. Lastly, it is investigated whether the obtained results also hold for different test sets. For BMA, different weighted combinations of the top 10 models are evaluated. Here it is investigated which inclusion rules result in the combinations with the highest prediction performance. Next, the selected combinations are weighted by the traditional model averaging criteria and the results are compared. Lastly, it is again investigated whether the obtained results hold for different test sets. Finally, the prediction performance of the BVS models is compared to the BMA combinations.

Figure 3: An overview of the analysed algorithms, selection criteria and their corresponding prediction performance criteria



2.8 Programming

The programming language used for the methodology as described in this paper is R. Additionally, the MCMC simulations and the corresponding models are estimated using Just Another Gibbs Sampler (JAGS). These two languages are connected by using the 'RJAGS' package in R. All other packages and functions are available in R. The parameters of the logistic regression, estimated with IRLS, are obtained with the 'glm' function of the 'stats' package, by choosing 'method = glm.fit'. The sampling of the posterior probability is done with the 'coda.sampling' function from the 'RJAGS' package. The predictions and prediction accuracy are obtained using the 'prediction' and 'performance' functions in the 'ROCR' package.

2.9 Assumptions

In this section, the assumptions that are made throughout this paper are described. For Bayesian inference, it is assumed that the observations of the dependent variable are independently distributed, conditional on the unknown parameters. This assumption is met as all observations are randomly sampled from the dataset. Additionally, assumptions are made about the distribution of the priors and likelihood as described above. The prior odds of the different models is assumed to be equal. If information about these probabilities was available a priori, this could be used to improve the effectiveness of the BVS. It is assumed that the sampling $\boldsymbol{\gamma}$ through SSVS results in an accurate representation of the posterior distribution

$f(\boldsymbol{\gamma}|\boldsymbol{y})$. Lastly, it is assume that this posterior distribution of $f(\boldsymbol{\gamma}|\boldsymbol{y})$ is a representation of the prediction performance of the different models.

For the logistic regression, we make assumptions about the analysed data set. As Bewick et al. (2005) describe, we assume that the dependent variable is binomial distributed, and the explanatory variables are not strongly correlated. As the variables CLTV and LTV have a correlation of 95%, we correct for this by subtracting the value of LTV from CLTV, effectively decreasing this correlation to 11%. Additionally, the observations are assumed to be independent and the data set should be sufficiently large.

3 Data

Table 1: The explanatory variables, their description, range limits, their label used in the coding, and the unit of measurement.

Variable	Description	Range	Code Label	Unit
Credit Score	A number summarizing the borrowers creditworthiness, prepared by a third party. Also known as Fico Score.	301-850	fico	#
Insurance	The percentage of loss coverage on the loan, at the time of Freddie Mac's purchase of the mortgage loan that a mortgage insurer is providing to cover losses incurred as a result of a default on the loan.	1-55	mi_pct	%
Units	Binary variable indicating whether the mortgage is a one-, or more-unit property.	0 = One 1 = More than one	cnt_units	#
CLTV	Combined Loan-To-Value (CLTV). In the case of a purchase mortgage loan, the ratio is obtained by dividing the original mortgage loan amount on the note date and any secondary mortgage loan amount disclosed by the seller, by the lesser of the mortgaged property's appraised value on the note date or its purchase price.	0-200	CLTV	%
DTI	Debt-To-Income (DTI). Disclosure of the debt to income ratio is based on (1) the sum of the borrower's monthly debt payments, including monthly housing expenses that incorporate the mortgage payment the borrower is making at the time of the delivery of the mortgage loan to Freddie Mac, divided by (2) the total monthly income used to underwrite the loan as of the date of the origination of the loan.	0-65	dti	%
UPB	Unpaid Principle Balance (UPB). The amount of UPB of the mortgage on the note date.	UPB >0	orig_upb	$
LTV	Loan-To-Value (LTV). In the case of a purchase mortgage loan, the ratio obtained by dividing the original mortgage loan amount on the note date by the lesser of the mortgaged property's appraised value on the note date or its purchase price.	6-105	ltv	%
Interest Rate	The note rate as indicated on the mortgage note.	0-100	int_rt Rate	%
Borrowers	A binary variable indicating whether there is only one, or more than one, borrower obliged to pay the mortgage note.	0 = One 1 = More than one	cnt_borr	#

This paper uses data from the Freddie Mac's Single Family Loan-Level data set. The full data set includes information on fully amortizing fixed-rate mortgages that Freddy Mac bought between 1999 and 2017. For each year, there is a random sample available, containing the information of 50.000 mortgage loans. The data set consists of a large amount of potential explanatory variables, from which $p = 9$ are selected. The default of a loan in this research is defined as a loan that is past due for three months or longer, and is converted to a binary variable indicating a default as 1 and a non-defaults as 0. The selected explanatory variables are all the available numerical variables in the Freddie Mac data set and three binary variables. Table 1 states a description and summary of these variables. The data set is not normalized or transformed, as this influences the economical interpretability of the results and is not required for the effective application of the methodology. We visually inspect the data set, however this will not lead to the deletion of any outliers. Additionally, the variable CLTV is corrected by the variable LTV to solve the appearing multicollinearity. Table 2 gives overview of the descriptive data statistics.

Table 2: Overview of the descriptive values of the explanatory variables of the train set.

	Credit Score	Insurance	Units	CLTV	DTI	UPB	LTV	Interest Rate	Borrowers
Min Value	300.0	0.0	0.0	0.0	1.0	21000.0	8.0	3.8	0.0
Mean	738.8	5.0	0.0	74.2	1.3	205987.0	73.0	5.3	0.5
Median	749.0	0.0	0.0	80.0	0.0	182000.0	78.0	5.3	1.0
Max Value	842.0	37.0	1.0	54.0	65.0	802000.0	100.0	7.8	1.0
Sd	52.6	10.6	0.2	16.2	4.7	111943.9	16.1	0.7	0.5

When modelling the PD, a distinction should be made between data that is analysed Through The Cycle (TTC) or at a Point In Time (PIT). Aguais et al. (2004) state that TTC data has a time span of at least five years, incorporating underlying cyclical patterns of recessions and expansions in the business cycle. This makes this type of data more suitable to be analysed for PD predictions with a longer time span. By taking a time span that includes the whole business cycle, its overall effects are cancelled out and predictions can be made about the long term average. Alternatively, PIT data has a time span of one year at the most, and therefore should be used to make short term predictions. These predictions are conditional on the position on the business cycle of that specific period in time. To be able to make unconditional predictions, information regarding this position is required. The contribution of this research lies not in creating a complex PD model that competes on TTC prediction performance with the best models present in the literature, but in investigating the added value of the the above described Bayesian methods in a new field of interest. Therefore, data from only four different years, at four different places in the business cycle and with equal time gaps between them, is analysed. That is, a pre-crisis sample of 2005, a during crisis sample of 2008, a post-crisis sample of 2011 and an outside-crisis sample of 2014. This can be seen as four different PIT data sets that are combined with the purpose of making the obtained results generally applicable in multiple stages of the business cycle.

This research starts with analyzing a small part of the full data set. The four yearly samples will each have 3000 randomly selected observation, resulting in a sample size of 12000. These are randomly split in a train and test sample, both including of $n = 6000$ observations. Only the PD within the first year is investigated. The analysis will contain the $p = 9$ described explanatory.

An explanatory variable that is generally highly informative for PD models is the variable arrear, indicating whether there has been an arrear in monthly payments in the past. However, this variable is not available for the model analysed in this paper, as we model the PD in the first year. Additionally, two random samples also consisting of $n = 6000$ observations are taken independently from the first sample. These are additionally used to test the consistency of the estimated out-of-sample prediction performance for different independent samples. Note that the amount of observed defaults is relatively low. The train set has 30 defaulting observations, whereas the three test sets have 30, 18, and 22 observed defaults respectively.

4 Results

4.1 Bayesian Variable Selection

In this section, the main results of BVS are displayed and commented on. First the model specifications are tuned to increase the efficiency of the SSVS algorithm. By experimenting with different values of c_i and $f(\boldsymbol{\gamma})$, the best data driven fit is investigated. Secondly, a list of the top 5 models with the highest posterior probability is presented and their regression coefficients and prediction performance are evaluated. Furthermore, the AIC and BIC of these models is analysed with the aim of investigating whether the effectiveness of BVS differs from these traditional variable selection criteria. Additionally, the more traditional variable selection algorithms FS, BE, and SS are used to obtain models that are used as a benchmark. Figure 4 gives a graphical overview of this section. Since we are comparing the model top 5, we set $h = 5$. Consequently the posterior probability of the 5 top models sums up to one. Note that for the readers convenience, the cells displaying the best prediction performance criteria value of a table are colored grey.

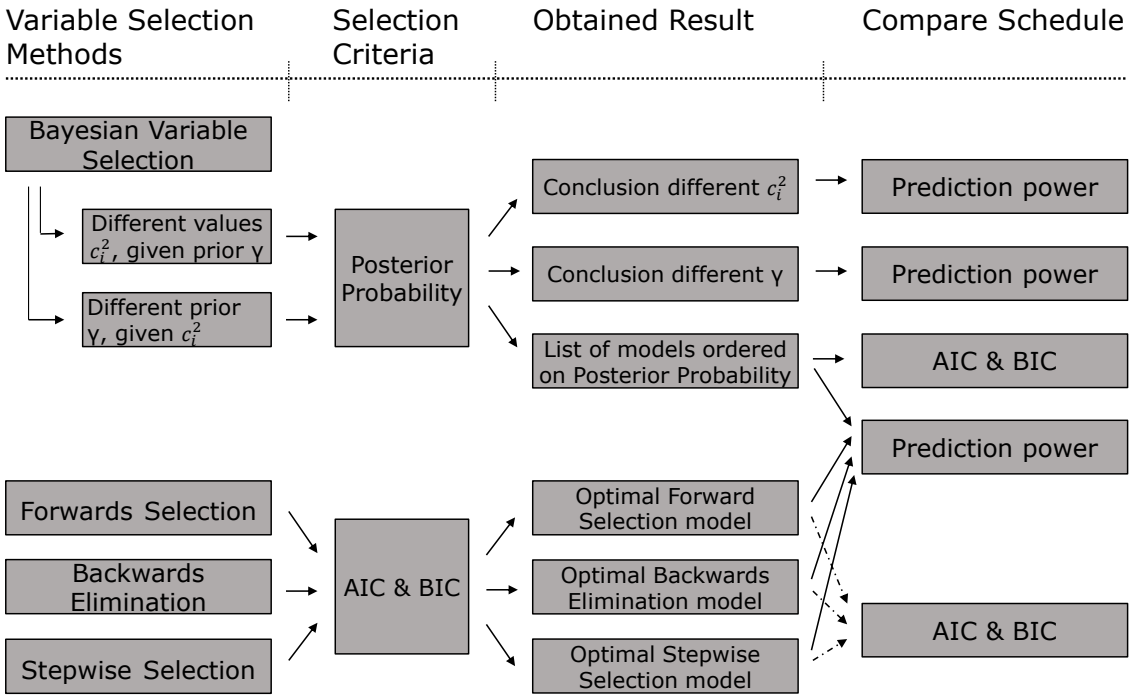


Figure 4: A graphical overview of the result section of BVS.

4.1.1 Tuning c_i and $f(\boldsymbol{\gamma})$

George & McCulloch (1993) describe different possibilities to obtain the optimal tuning values of c_i and $f(\boldsymbol{\gamma})$. Overall, it is stated that these potential solutions should not be seen as hard and fast rules, and c_i should be seen as a tuning constant that calibrates the available information about $f(\boldsymbol{\gamma})$. Furthermore, for $f(\boldsymbol{\gamma})$, the priors used in this paper are recommended, as they have proven to work well when used for practical application. O'Hara et al. (2009) suggest that data-driven priors should be used for the tuning variables with the goal of improving the mixing of the MCMC chains. This paper will implement this data driven approach, by exploring the effect of changes in the values c_i and $f(\boldsymbol{\gamma})$ separately, while keeping the other constant.

c_i can be interpreted as the prior odds of excluding x_i when β_i is close to zero, hence it defines the threshold between including or excluding an explanatory variable. Note that the relation between c_i and this threshold is not linear, as the threshold only slightly decreases when c_i increases. As shown in Equation 14, the effect of c_i depends on its size relative to τ, which is set at $\tau = \frac{1}{100}$ in this paper. c_i is defined for values larger than 0, and the initial value of c_i is set at $30,000$, which is roughly three times the mean of the standard deviation of the explanatory variables. p_i can be interpreted as the prior odds of including a explanatory variable into the model. Consequently, p_i is a probability, and is defined for values between 0 and 1. For $f(\boldsymbol{\gamma})$, we define the prior as described by Equation 20, starting with $f(\boldsymbol{\gamma}) = p_i^{-2}$ which indicates a uniform indifferent prior that states $P(\gamma_i = 1) = p_i = 0.5$, indicating that the probability of including a variable is equal to not including a variable. In this paper, both p_i

and c_i are assumed to be constant within an observation, that is $c_i = c$ and $p_i = p$. Finally, the values $c = 500.000$ and $p = 0.9$ are selected.

Table 3: Top 5 models with the highest posterior probability for different values for c, given $P(\gamma_i = 1) = 0.90$, and their corresponding prediction performance criteria. Here M_1 implies the model with highest posterior probability. PosProb = Posterior probability, FP60 = amount of FP observations given a TRP rate of 60%, FP90 = amount of FP observations given a TRP rate of 90%, FP50-100 = average amount of FP observations in the TPR range of $50\% - 100\%$, Z-value based on a TPR of 100%. K stands for thousand, and M stands for million.

	M_l	Top 5 models	PosProb	FN60	FN90	FN50-100	AUC	Z-value
$c = 5$	M_1	Credit Score, Insurance, Units, CLTV, DTI, UPB, LTV, Interest Rate, Borrowers	0.497	429	3073	1531	0.867	106.666
	M_2	Insurance, Units, CLTV, DTI, UPB, LTV, Interest Rate, Borrowers	0.127	425	3178	1765	0.844	145.413
	M_3	Credit Score, Insurance, Units, CLTV, DTI, UPB, Interest Rate, Borrowers	0.127	439	3045	1536	0.867	106.388
	M_4	Credit Score, Insurance, Units, CLTV, DTI, LTV, Interest Rate, Borrowers	0.126	279	2505	1280	0.890	127.033
	M_5	Credit Score, Units, CLTV, DTI, UPB, LTV, Interest Rate, Borrowers	0.122	453	2309	1377	0.879	108.152
$c = 500$	M_1	Insurance, Units, CLTV, Interest Rate, Borrowers	0.263	327	2573	1574	0.865	116.732
	M_2	Insurance, Units, Interest Rate, Borrowers	0.262	328	2591	1589	0.863	117.852
	M_3	Units, Interest Rate, Borrowers	0.185	344	3047	1547	0.870	100.021
	M_4	Units, CLTV, Interest Rate, Borrowers	0.154	330	2957	1503	0.874	98.179
	M_5	Insurance, Units, CLTV, LTV, Interest Rate, Borrowers	0.137	330	2454	1543	0.865	119.957
$c = 50K$	M_1	Insurance, Units, Interest Rate, Borrowers	0.362	328	2591	1589	0.863	117.852
	M_2	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.272	283	2460	1283	0.890	126.119
	M_3	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.174	288	2488	1291	0.889	126.655
	M_4	Insurance, Units, CLTV, Interest Rate, Borrowers	0.129	327	2573	1574	0.865	116.732
	M_5	Credit Score, Insurance, Units, DTI, Interest Rate, Borrowers	0.063	282	2454	1279	0.890	126.119
$c = 500K$ & c = 5M	M_1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.395	283	2460	1283	0.890	126.119
	M_2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.278	288	2488	1291	0.889	126.655
	M_3	Credit Score, Insurance, Units, LTV, Interest Rate, Borrowers	0.142	280	2479	1274	0.890	125.959
	M_4	Credit Score, Insurance, Units, CLTV, LTV, Interest Rate, Borrowers	0.093	280	2503	1282	0.890	126.871
	M_5	Credit Score, Insurance, Units, DTI, Interest Rate, Borrowers	0.093	282	2454	1279	0.890	126.119
$c = 50M$	M_1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.404	283	2460	1283	0.890	126.119
	M_2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.240	288	2488	1291	0.889	126.655
	M_3	Credit Score, Insurance, Units, LTV, Interest Rate, Borrowers	0.171	280	2479	1274	0.890	125.959
	M_4	Credit Score, Insurance, Units, CLTV, LTV, Interest Rate, Borrowers	0.114	280	2503	1282	0.890	126.871
	M_5	Credit Score, Insurance, Units, CLTV, DTI, Interest Rate, Borrowers	0.071	282	2482	1286	0.889	126.763
$c = 500M$	M_1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.615	283	2460	1283	0.890	126.119
	M_2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.165	288	2488	1291	0.889	126.655
	M_3	Credit Score, Insurance, Units, DTI, Interest Rate, Borrowers	0.089	282	2454	1279	0.890	126.119
	M_4	Credit Score, Insurance, Interest Rate, Borrowers	0.074	289	2463	1287	0.889	126.763
	M_5	Credit Score, Insurance, CLTV, Interest Rate	0.057	394	1741	1376	0.879	196.345

Table 3 displays the effect of different values for c on the obtained top 5 models in terms of posterior probability, and shows the corresponding values of their prediction performance criteria. Here, p is set at 0.9. Setting c low results in a low threshold for including variables, and results in including many variables into the model. This can be seen in the top models for $c = 5$, that incorporate a relatively large amount of explanatory variables which results in low prediction performance for almost every model. When evaluating c in the range of 500 to $500K$, the obtained optimal models, and the number of included variables, are constantly changing. Here, the prediction performance on average increases as the value of c increases. For c in the range of $500K$ to $5M$, the optimal models do not change, and the only differences is the posterior probability. The results of $c = 5m$ are presented in the appendix. For values of c that are higher than $5M$, the prediction performance becomes less good. Moreover, it can be seen that models with high and low values for c have relatively higher posterior probabilities for their best models. This can result in models that are selected because of the fit of their explanatory variables with the model set-up, rather than because their explanatory variables are more likely to describe the underlying data. Especially for the analysis of model averaging this has large consequences, as disproportional values of the posterior probability directly affect the BMA weights and consequently the obtained results. Consequently $c = 500.000$ is set for the remainder of this paper, as it has the highest prediction performance criteria values and non-disproportional values for the posterior probability.

Table 4: Top 5 models ordered on their highest posterior probability for different values for p, given $c = 500,000$, and their corresponding prediction performance criteria. Here M_1 implies the model with highest posterior probability. PosProb = Posterior probability, FP60 = amount of FP observations given a TRP rate of 60%, FP90 = amount of FP observations given a TRP rate of 90%, FP50-100 = average amount of FP observations in the TPR range of $50\% - 100\%$, Z-value based on a TPR of 100%.

	M_l	Top 5 models	PosProb	FN60	FN90	FN50-100	AUC	Z-value
$p = 0.95$	M_1	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.302	288	2488	1291	0.889	126.66
	M_2	Credit Score, Insurance, Units, CLTV, LTV, Interest Rate, Borrowers	0.236	280	2503	1282	0.890	126.87
	M_3	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.229	283	2460	1283	0.890	126.12
	M_4	Credit Score, Insurance, Units, LTV, Interest Rate, Borrowers	0.127	280	2479	1274	0.890	125.96
	M_5	Credit Score, Insurance, Units, CLTV, DTI, LTV, Interest Rate, Borrowers	0.106	279	2505	1280	0.890	127.03
$p = 0.90$	M_1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.395	283	2460	1283	0.890	126.119
	M_2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.278	288	2488	1291	0.889	126.655
	M_3	Credit Score, Insurance, Units, LTV, Interest Rate, Borrowers	0.142	280	2479	1274	0.890	125.959
	M_4	Credit Score, Insurance, Units, CLTV, LTV, Interest Rate, Borrowers	0.093	280	2503	1282	0.890	126.871
	M_5	Credit Score, Insurance, Units, DTI, Interest Rate, Borrowers	0.093	282	2454	1279	0.890	126.119
$p = 0.85$	M_1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.506	283	2460	1283	0.890	126.12
	M_2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.220	288	2488	1291	0.889	126.66
	M_3	Credit Score, Insurance, Units, LTV, Interest Rate, Borrowers	0.103	280	2479	1274	0.890	125.96
	M_4	Insurance, Units, Interest Rate, Borrowers	0.089	328	2591	1589	0.863	117.85
	M_5	Credit Score, Insurance, Units, DTI, Interest Rate, Borrowers	0.081	282	2454	1279	0.890	126.12
$p = 0.80$	M_1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.487	283	2460	1283	0.890	126.12
	M_2	Credit Score, Insurance, Interest Rate, Borrowers	0.211	289	2463	1287	0.889	126.76
	M_3	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.134	288	2488	1291	0.889	126.66
	M_4	Insurance, Units, Interest Rate, Borrowers	0.111	328	2591	1589	0.863	117.85
	M_5	Credit Score, Insurance, Units, LTV, Interest Rate, Borrowers	0.059	280	2479	1274	0.890	125.96
$p = 0.50$	M_1	Interest Rate, Borrowers	0.473	355	3075	1565	0.869	100.71
	M_2	Insurance, Interest Rate, Borrowers	0.208	329	2604	1597	0.863	118.66
	M_3	Interest Rate	0.132	386	3147	1804	0.855	134.71
	M_4	Insurance, Units, Interest Rate, Borrowers	0.095	328	2591	1589	0.863	117.85
	M_5	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.092	283	2460	1283	0.890	126.12

Table 4 displays the effect of different values for p on the top 5 models in terms of posterior probability, and shows their corresponding prediction performance. Here c is set at $500,000$. As for extreme values, setting $p = 0$ includes zero variables into the model, resulting in a model with

only a constant and very little prediction performance. On the contrary, setting $p = 1.0$ results in a model that includes all explanatory variables. The vague prior $p = 0.5$ results in relatively small models, where the model with the highest probability only has two variables and low prediction performance. Setting the probability of including a variable higher results in models with a larger amount of explanatory variables. When focusing on all prediction performance criteria except for the Z-value, setting $p = 0.9$ results in the best performing models. Hence, this is set as the prior value of p for the remainder of this paper.

4.1.2 The optimal Bayesian Variable Selection Models

Table 5: Top 5 models obtained through BVS for tuning values $p = 0.9$ and $c = 500,000$.

	Model	PosProb
Model 1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.395
Model 2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.278
Model 3	Credit Score, Insurance, Units, LTV, Interest Rate, Borrowers	0.142
Model 4	Credit Score, Insurance, Units, CLTV, LTV, Interest Rate, Borrowers	0.093
Model 5	Credit Score, Insurance, Units, DTI, Interest Rate, Borrowers	0.093

Given the above described tuning values, Table 5 gives a list of the top 5 BVS models based on highest posterior probability and displays the included variables. As we are investigating the effectiveness and accuracy of BVS, we separately analyse the top 5 models that BVS labels as most likely. The relative likeliness between the models can be described with Bayes factors, as shown in Section 2.3. Consequently, the Bayes factor of Model 1 to Model 2 can be estimated by $\frac{0.395}{0.278} = 1.421$. This indicates that, given that either Model 1 or Model 2 is correct, the probability that model 1 is correct is estimated by $\frac{1.421}{1+1.421} = 0.0.587$. Also note that Model 4 and Model 5 are equally likely. All selected models include an intercept, and the explanatory variables Credit Score, Insurance, Units, Interest Rate, and Borrowers. The difference in the models lies in the inclusion of CLTV, DTI and LTV. UPB is not included in any of the models.

Table 6: Regression coefficients of the top 5 models with the highest posterior probability given $p = 0.90$ and $c = 500,000$, and the FS and BE SS models. For the explanatory variables included in the models, the regression coefficients are stated with their standard errors in parenthesis. * = Coefficient significantly differs from 0, for a 95% significance level.

	Intercept	Credit Score	Insurance	Units	CLTV	DTI	UPB	LTV	Interest Rate	Borrowers
Model 1	-7.072*	-0.010*	0.054*	-0.160					1.509*	-1.115*
	(2.930)	(0.003)	(0.014)	(1.044)					(0.288)	(0.443)
Model 2	7.076*	-0.010*	0.054*	-0.156	0.008				1.510*	-1.116*
	(2.931)	(0.003)	(0.014)	(1.045)	(0.058)				(0.289)	(0.443)
Model 3	-7.269*	-0.010*	0.051*	-0.155				0.003	1.508*	-1.115*
	(3.330)	(0.003)	(0.024)	(1.045)				(0.024)	(0.288)	(0.443)
Model 4	-7.273*	-0.010*	0.052*	-0.151	0.008			0.003	1.508*	-1.115*
	(3.330)	(0.003)	(0.024)	(1.046)	(0.059)			(0.024)	(0.289)	(0.443)
Model 5	-7.104*	-0.010*	0.054*	-0.162		0.001			1.507*	-1.114*
	(2.996)	(0.003)	(0.014)	(1.045)		(0.016)			(0.290)	(0.444)
FS	-8.223*	-0.012*	0.057*	-0.846	-0.012	-0.007	6.320E-06*	0.003	1.664*	-1.317*
	(3.552)	(0.003)	(0.025)	(1.104)	(0.062)	(0.017)	(1.759E-06)	(0.025)	(0.297)	(0.451)
BE & SS	-7.994*	-0.011*	0.060*				5.726E-06*		1.614*	-1.262*
	(2.991)	(0.003)	(0.014)				(1.605E-06)		(0.291)	(0.447)

The first 5 models in Table 6 show the regression coefficients of the top BVS models, calculated with IRLS. The regression estimates as calculated through BVS are presented in Equation 17 in the Appendix. They do not significantly differ from the IRLS results. Except for Units, the coefficients of all variables that are included in all 5 top models, significantly differ from 0 at a 95% significance level. Hence, depending on the sign of the coefficient, we can say that there is at least a 95% probability that these variables have a positive or negative influence on the PD of the mortgage portfolios. When the variables CLTV or LTV are included, they are insignificant and only slightly change the coefficients of the other included variables. Including DTI also gives an insignificant regression coefficient, however this has more impact on the regression coefficients of the other included variables. The variable UPB is never included, which indicates that this variable has a negative impact on the posterior probability of a model. Hence, according to BVS, this variable has a negative impact on the prediction performance of a model.

For the interpretation of the coefficients, we look at its relation to the odds ratio of the PD as described in Equation 3. Recall that the odds ratio of the PD is described by the probability of a default, divided by the probability of a non-default. Hence the regression coefficient represent the changes in the log odds ratio of the PD. When assuming that the other variables stay constant, the change in the odds ratio resulting from an increase of 1 unit in variable x_i can be described by e^{β_i}. Here β_i indicates the regression coefficient as presented in the Table 6. The odds ratio is converted to the actual change in PD as described in Equation 2.

We analyse the binary variable Borrowers of model 1 as an example. The regression coefficient is -1.115. Consequently, the odds ratio of having more than one borrower versus having only one borrower is $e^{-1.115} = 0.328$. Hence the average difference in the PD when having more than one borrower is $\frac{0,328}{1+0.328} - \frac{1}{1+0.328} = -0.506$, all else equal. This negative influence of multiple borrowers on the PD can be explained by the fact that, when one borrower is not able to pay the mortgage, the other borrowers are responsible. As a second example we show the effect of the continuous variable Interest Rate of Model 1. Note that the Interest Rate values in the data set range from $3,75\%$ until $8,38\%$. An increase of 1% in the Interest Rate, results in an average increase of $e^{1,509} = 4.522$ in the odds of defaulting, independent of the base Interest Rate and assuming all else equal. The actual difference in PD depends on the base value of the

Interest Rate. This positive relation between Interest Rate and PD is a logical consequence of a higher risk premium when the mortgage loan has a higher PD.

It follows that significantly positive regression coefficients have a positive impact on the PD. Hence, we can say that the variable Insurance has a positive effect on the PD of a variable. This can be explained by adverse selection, as it is more profitable for loans with a higher PD to buy an insurance against a default. Furthermore, the variable Credit Score describes the borrowers creditworthiness, hence the negative relation with the PD is as expected. The variable ' UPB' is not included in the first five models. However the positive relation as described by the last two models can be explained as a larger sum of remaining unpaid principle balance takes a longer time, or larger monthly payments, to pay back. Both cases logically increase the PD. The signs of the remaining variables do not significantly differ from zero, hence no conclusions about their impact can be made.

Table 7: Top 5 models with the highest posterior probability given $P = 0.90$ and $c_i = 500,000$, and their corresponding prediction performance criteria. The included explanatory variables, AIC and BIC are calculated by the training set. The prediction performance criteria are calculated with test set 1. PosProb = Posterior probability, FP60 = amount of FP observations given a TRP rate of 60%, FP90 = amount of FP observations given a TRP rate of 90%, FP50-100 = average amount of FP observations in the TPR range of 50% − 100%, Z-value based on a TPR of 100%.

	Model	AIC	BIC	FN60	FN90	FN50-100	AUC	Z-value
BVS model 1	Credit Score, Insurance, Units, Interest Rate, Borrowers	291.170	331.367	283	2460	1283	0.890	126.119
BVS model 2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	293.152	340.049	288	2488	1291	0.889	126.655
BVS model 3	Credit Score, Insurance, Units, LTV, Interest Rate, Borrowers	293.154	340.050	280	2479	1274	0.890	125.959
BVS model 4	Credit Score, Insurance, Units, CLTV, LTV, Interest Rate, Borrowers	295.136	348.732	280	2503	1282	0.890	126.871
BVS model 5	Credit Score, Insurance, Units, DTI, Interest Rate, Borrowers	293.167	340.064	282	2454	1279	0.890	126.119
FS model	Credit Score, Insurance, Units,CLTV, DTI, UPB, LTV, Interest Rate, Borrowers	287.398	354.393	429	3073	1531	0.867	106.666
BE & SS model	Credit Score, Insurance, UPB, Interest Rate, Borrowers	280.327	320.524	442	2952	1506	0.870	112.867

In Table 7, the first 5 models presented are again the top BVS models with the highest posterior probability, The FP60 and FP90 criteria indicate the number of FP, given a TPR of 60% and 90%, e.g., the best obtained FP60 rate of 280 indicates that in order to correctly predict 60% of the defaults, 280 non-defaulting portfolios are falsely predicted as a default. This corresponds to $\frac{280}{6000-30} = 4,71\%$ wrongly predicted non-defaulting portfolios. Here 30 indicates the total amount of defaults in the test data set of $n = 6000$ observations. Note however, that these criteria only indicate the prediction performance of the model at the specific TPR rates. An overall more informative criteria is FN50-100, which indicates the average rounded number of FP within the TRP range of $50\% - 100\%$. A lower value here indicates that on average less FP are needed for a model, for the indicated TPR range. The AUC criteria indicates the models ability to discriminate between default and non-defaults. As this criteria describes the whole TRP range, it is considered informative. Lastly, the Z-value indicates the models prediction accuracy at a TPR rate of 100%, which is corrected by the variation.

When looking at the order of the models, BVS is not able to order them strictly from the highest to the lowest prediction performance. However, these are the top 5 models from a

set of $2^9 = 512$ models, hence roughly the top 1% models. When comparing these 5 models with all other models evaluated in this paper, four of them are among the overall top 5 models with the highest prediction performance. This indicates that BVS succeeds in attaching high probabilities to models with a relatively high prediction performance. When looking at the traditional variable selecting criteria AIC and BIC for these 5 models, Model 4 and Model 5 would have been substituted, which results in a better order of the prediction power. Hence given the top 5 as calculated by BVS, the traditional criteria result a better ordered list in this case.

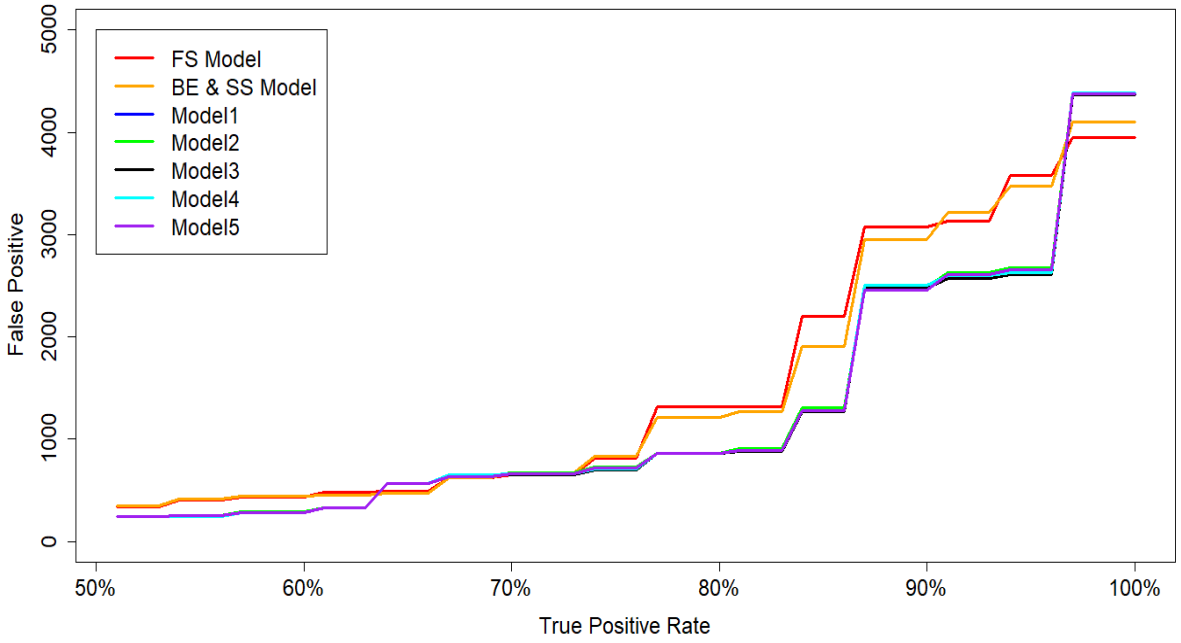


Figure 5: Amount of FP given the TPR plotted in the range of $50\% - 100\%$. Based on test set 1.

4.1.3 Comparison to the Traditional Benchmark Models

The traditional algorithms FS, BE, and SS are used to create benchmark models. Both optimizing criteria AIC and BIC result in identical models for all three algorithms. Furthermore BE and SS both obtain the same model. The model obtained by FS and the model obtained by both BE and SS are presented as the last two models in Table 6 and Table 7. The AIC value of the benchmark model is lower compared to the BVS models. However, when looking at the BIC, the value of the FS model is higher compared to the other models. This is an unexpected result, as the algorithm explicitly filters models that have a high value for this criteria. This indicates that the FS algorithm stopped at a local optimum.

In the FS model, all potential variables are included. The variables included in Model BE & SS are also different from the BVS models. Consequently, the estimated regression coefficients and hence the corresponding predictions are also different. When we make the comparison between the prediction performance of the BVS models and the benchmark models, the BVS models have higher criteria values, except for the Z-value. The reason for this is visualized in Figure 5, which gives a graphical representation of the required FP plotted against the TPR

values. It shows that the prediction performance of the BVS models in general is better than the traditional models. However, at a TPR of 100% the traditional models result in a lower amount of FP. This indicates that, when restricting the model to predicting all actual defaults, the BVS models do not have the highest prediction performance. Is is concluded that the BVS models on average outperform the benchmark models.

4.1.4 Multiple Test Sets

In order to prevent accepting specific results as general conclusions, two extra test samples of $n = 6000$ are taken from the remainder of the full data set, as described in Section 3. Tables and figures similar to Table 7 and Figure 5, but estimated using the data of the other two test sets are presented in the appendix. Note that the posterior probability, AIC and BIC are calculated using the training set, and hence do not change when we estimate using another test set. Furthermore, for these extra test sets, it is concluded that most of the above described findings are considered to be general. For every test set, the BVS models outperform the benchmark on average, but the benchmark has a lower Z-value. This signals that BVS focuses on the whole TPR range, whereas the traditional criteria focus on models that predict all of the actual defaults. Model 4 has the highest AIC and BIC values, but the best prediction performance for the other two test sets. Hence, here BVS is better in ordering the prediction performance of the models within the top 5. Another observed difference is that for both extra sets, the FP60 criteria is lower for the traditional selected models. As these results were not observed for test set 1, no general conclusion can be drawn.

In general, it is concluded that, for high TPR's, the BVS significantly outperform the benchmark. However, when requiring a TPR of 100%, the benchmark models have a higher prediction performance.

4.2 Bayesian Model Averaging

In this section, the main BMA results are reported and commented on. For BVS, the goal is to eventually select one optimal model, whereas BMA aims to combine multiple models with a high probability. Therefore, in order to more thoroughly investigate BMA and its potential model combinations, both the top 10 and top 5 models is analysed. Table 8 displays the top 10 models ordered on posterior probability. Different combinations of these models are evaluated on their prediction performance. These are both simple combinations and combinations made according to the rules of thumb as described in Section 1. For BMA, the different models are weighted by the relative size of their posterior probability. Additionally, the results of the same model combinations, but weighted by their AIC, BIC or using equal weights are used as a benchmark. Lastly, the results of the two additional test sets are analysed with the goal of making the application of the conclusions more general. Figure 6 gives a detailed overview of this part of the result section. Furthermore, in this section it is investigated which rules result in the combinations with the best prediction performance.

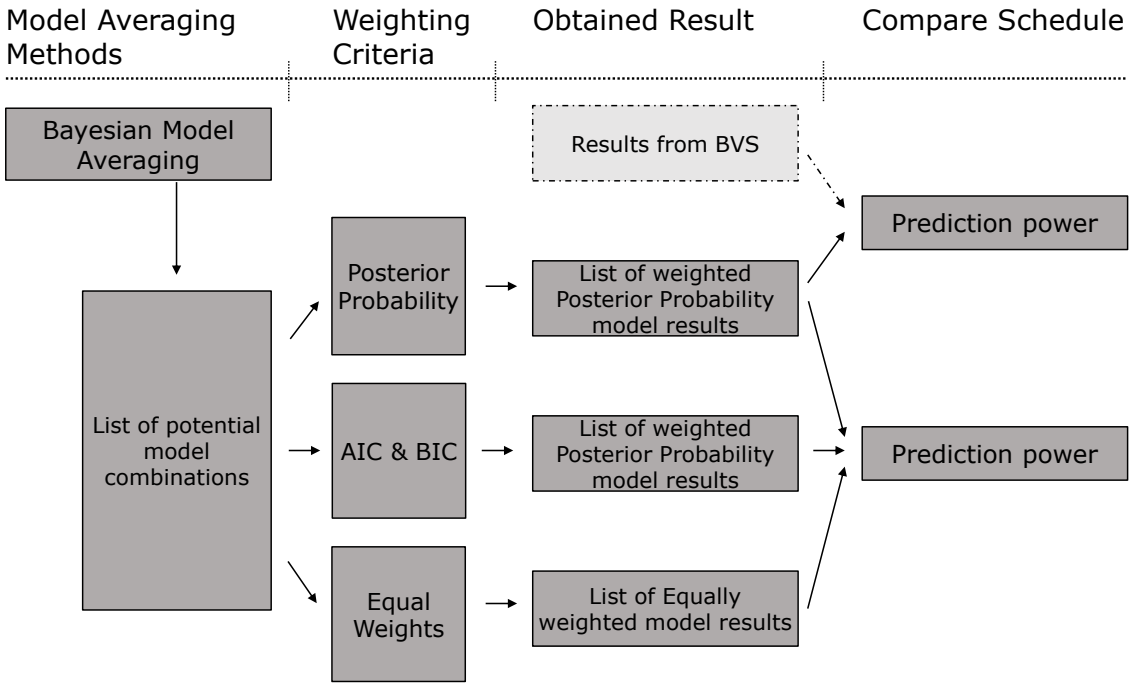


Figure 6: An overview of the result section of Bayesian Model Averaging, including a comparison schedule.

Table 8: Prediction performance of the top 10 models with the highest posterior probability, as calculated with BVS. Estimated with test set 1.

	Variables	PosProb	AIC	BIC	FP60	FP90	FP50-100	AUC	Z-score
Model 1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.328	291.170	331.367	283	2460	1283	0.890	126.119
Model 2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.231	293.152	340.049	288	2488	1291	0.889	126.655
Model 3	Credit Score, Insurance, Units, LTV, Interest Rate, Borrowers	0.118	293.154	340.050	280	2479	1274	0.890	125.959
Model 4	Credit Score, Insurance, Units, CLTV, LTV, Interest Rate, Borrowers	0.077	295.136	348.732	280	2503	1282	0.890	126.871
Model 5	Credit Score, Insurance, Units, DTI, Interest Rate, Borrowers	0.077	293.167	340.064	282	2454	1279	0.890	126.119
Model 6	Credit Score, Insurance, Units, CLTV, DTI, Interest Rate, Borrowers	0.054	295.149	348.746	282	2482	1286	0.889	126.763
Model 7	Credit Score, Insurance, Interest Rate, Borrowers	0.048	289.194	322.692	289	2463	1287	0.889	126.763
Model 8	Insurance, Units, Interest Rate, Borrowers	0.026	300.384	333.882	328	2591	1589	0.863	117.852
Model 9	Credit Score, Insurance, CLTV, Interest Rate, Borrowers	0.025	291.175	331.372	290	2490	1295	0.889	127.413
Model 10	Credit Score, Insurance, Units, DTI, LTV, Interest Rate, Borrowers	0.016	295.151	348.747	278	2477	1271	0.891	126.280

4.2.1 Analysis of Model Combinations

Table 9: Prediction performance of analysed different simple combinations of models, weighted with BMA. The combined models are described in Table 8.

BMA simple combinations	FP60	FP90	FP50-100	AUC	Z-score
Model 1 & 2	287	2476	1287	0.889	126.333
Model 1, 2 & 3	282	2479	1285	0.890	126.173
Model 1, 2, 3 & 4	281	2480	1284	0.890	126.226
Model 1, 2, 3, 4 & 5	281	2480	1284	0.890	126.280
Model 1, 2, 3, 4, 5 & 6	281	2480	1284	0.890	126.280
Model 1, 2, 3, 4, 5, 6 & 7	281	2479	1284	0.890	126.280
Model 1, 2, 3, 4, 5, 6, 7 & 8	279	2501	1288	0.889	126.066
Model 1, 2, 3, 4, 5, 6, 7, 8 & 9	279	2501	1288	0.889	126.119
Model 1, 2, 3, 4, 5, 6, 7, 8 , 9 & 10	279	2500	1287	0.889	126.066

Table 9 present different simple BMA combinations of the top models calculated by BVS, ordered on their posterior probability. We start with the two models with the highest posterior probability and sequentially add the next best model for every new combination, until a combination of all 10 top models is obtained. The first three BMA combinations in Table 10 are the same simple combinations of the first two, first three and all ten models. The remaining combinations are selected using the described inclusion rules, applied to the top 5 and top 10 models respectively. That is, the combinations that are in the blue cells, are made out of models that have a high prediction performance when analysed individually. The yellow combinations are created by the rule that excludes a model from the combination when a subset of this model is available with a better prediction performance. The green combinations are made using both inclusion rules.

Table 10: Prediction performance of model combinations created by applying the inclusion rules, and weighted with BMA. The blue cells indicate combinations that are made by the rule that includes models with a high prediction performance. The yellow combinations are made by the rule that excludes a model from the combination when a subset of this model is available that has better prediction performance. The green combinations are made using both inclusion rules. The combined models are described in Table 8.

BMA combinations	FP60	FP90	FP50-100	AUC	Z-score
Model 1 & 2	287	2476	1287	0.889	126.333
Model 1, 2 & 3	282	2479	1285	0.890	126.173
All models	279	2500	1287	0.889	126.066
Model 1, 3, 4 & 5	281	2472	1280	0.890	126.119
Model 1, 3, 4, 5 & 10	281	2472	1280	0.890	126.119
Model 1, 3 & 5	281	2465	1280	0.890	125.960
Model 1, 3, 5, 7, 8, 9 & 10	274	2508	1287	0.889	125.641
Model 3 & 5	280	2470	1276	0.890	126.173
Model 1, 3, 5 & 10	281	2465	1280	0.890	126.066

For the simple combinations, the results indicate that middle sized combinations of four to seven models, with larger posterior probability, outperform small and large combinations. Additionally, Table 10 shows that the combinations that are created by applying both rules outperform all other combinations including the simple ones. This is concluded as they have the best values for the criteria that are based on the strongest prediction performance criteria FP50-100 and AUC. Note that the combinations created by excluding the models that have better

subsets show strong TPR specific criteria values for FN60, FN90, and the Z-value. However, overall they have a lower prediction performance. Because of the large diversity in the number of combined models in the best obtained combinations, no general conclusions can be drawn about the optimal number of combined models. Furthermore, the general application of the results is investigated in Section 4.2.3.

4.2.2 Comparison to the Traditional Benchmark Combinations

Table 11: Prediction performance of analysed different combinations of models, weighted equally. Based on test set 1.

AIC combinations	FP60	FP90	FP50-100	AUC	Z-score
Model 1 & 2	287	2478	1288	0.889	126.494
Model 1, 2 & 3	281	2479	1283	0.890	126.226
All models	261	2595	1299	0.888	124.329
Model 1, 3, 4 & 5	281	2479	1280	0.890	126.280
Model 1, 3, 4, 5 & 10	280	2478	1278	0.890	126.280
Model 1, 3 & 5	281	2465	1279	0.890	126.066
Model 1, 3, 5, 7, 8, 9 & 10	255	2581	1301	0.888	123.145
Model 3 & 5	280	2467	1276	0.890	126.119
Model 1, 3, 5 & 10	280	2467	1276	0.890	126.119

In this section, the above described results of the BMA combinations are compared to the benchmark combinations weighted by the traditional weights. Table 11 displays the prediction performance of the selected combinations weighted equally. The results obtained using AIC and BIC as weight are only marginally different from the equally weighted results. This is a logical consequence from the fact that the AIC and BIC values of the different models are large and only slightly different, hence using them as a weight is comparable to using equal weights. Consequently, only the AIC results are presented and discussed. It is concluded that the results of the optimal BMA combinations are almost similar to the results of the traditional weighted models. Specifically, when looking at the average performance of the combinations, the differences between the methods are neglectable. However, the differences become larger for combinations that include both models located in the upper and lower area of the top 10. This is as expected, as the larger difference in posterior probability of these models make the weights differ from the traditional benchmark. The finding that the combinations created by applying both rules perform better on average, is also applicable for the traditionally weighted benchmark combinations. The differences between the results obtained using the other traditional weights are small. As this difference is small, the results of the other two test sets is investigated in the next section.

4.2.3 Multiple Test Sets

In order to obtain more generally applicable conclusions, the prediction performance of the selected combinations is analysed for the other two test sets. For the simple combinations, the results of the other two sets are displayed in the appendix. The above described results are in favor of the middle sized models combinations. The other two sets show results in

favor of small to medium sized sets. Overall, models with a high ranking in the top 10 have a higher prediction performance. Combining models that individually have a high prediction performance results in a high prediction performance. Hence, in general it is concluded that small to medium combinations of models with a high posterior probability outperform larger model combinations. No evidence is discovered for an optimal amount of included models. Table 12 gives an overview of the average criteria values of differently obtained combinations.

Table 12: Average values of the criteria values for different combinations and models. (5) states that the top 5 models are used for the combinations, where (10) indicates the top 10 model are used. 'Same set 2 rules' indicates the average criteria values over the three sets, for the model combination obtained by applying both rules to the results of that same set. 'Other set 2 rules' indicates that the model combinations are made with another set. 'Train set 2 rules' indicates that the model combinations are made with the train set

Table Averages	FP60	FP90	FP50-100	AUC	Z-score
Simple combination (5)	719	2647	1620	0.857	207.454
Simple combination (10)	719	2638	1617	0.857	207.904
Same set 2 rules (5)	709	2653	1607	0.858	202.820
Same set 2 rules (10)	687	2624	1609	0.857	207.383
Other set 2 rules (5)	710	2653	1611	0.857	203.683
Other set 2 rules (10)	720	2647	1621	0.857	207.927
Train set 2 rules (5)	718	2636	1618	0.857	208.105
Train set 2 rules (10)	693	2634	1618	0.857	209.243
Individual models (5)	712	2642	1615	0.857	206.466
Individual models (10)	705	2718	1664	0.853	209.054

4.2.3.1 Analysis of Additional Model Combinations

Table 13: Prediction performance of different BMA combinations estimated with test set 1.The different combinations are based on applying the model inclusion rules to the top 10 models estimated with test set 1 (orange), 2 (light blue) 3 (red).

BMA combinations	FP60	FP90	FP50-100	AUC	Z-score
Model 1 & 2	287	2476	1287	0.889	126.333
Model 1, 2 & 3	282	2479	1285	0.890	126.173
All models	279	2500	1287	0.889	126.066
Combinations based on set 1					
Model 1, 3, 4 & 5	281	2472	1280	0.890	126.119
Model 1, 3, 4, 5 & 10	281	2472	1280	0.890	126.119
Model 1, 3 & 5	281	2465	1280	0.890	125.960
Model 1, 3, 5, 7, 8, 9 & 10	274	2508	1287	0.889	125.641
Model 3 & 5	280	2470	1276	0.890	126.173
Model 1, 3, 5 & 10	281	2465	1280	0.890	126.066
Combinations based on set 2					
Model 2, 3 & 4	281	2491	1284	0.890	126.602
Model 2, 3, 4, 6, 9 & 10	281	2492	1284	0.890	126.655
Model 1, 2, 3 & 4	281	2480	1284	0.890	126.226
Model 1, 2, 3, 4, 6, 7, 8 & 9	279	2505	1289	0.889	126.119
Model 3 & 4	280	2491	1278	0.890	126.280
Model 2, 3, 4, 6 & 9	281	2493	1285	0.890	126.602
Combinations based on set 3					
Model 2, 3 & 4	281	2491	1284	0.890	126.602
Model 2, 4, 7 & 9	285	2491	1289	0.889	126.709
Model 1, 2, 3 & 4	281	2480	1284	0.890	126.226
Model 7, 8 & 9	273	2580	1323	0.886	121.482
Model 3 & 4	280	2491	1278	0.890	126.280
Model 7 & 9	290	2474	1290	0.889	127.141

By applying the above described inclusion rules similarly to the other two test sets, new combinations are obtained. Next, all different combinations are estimated using the three test sets. Table 13 displays the results of all different combinations, estimated with test set 1. The first nine combinations are similar to the ones described above, but stated again to make the comparison to the other test sets. Here, the orange colored cells indicate the same combinations based on test set 1. The light blue cells indicate the combinations based on test set 2, and the red cells show the combinations based on test set 3. The results obtained when estimating these combinations with the other two test sets are displayed in the appendix. For the most informative criteria FN50-100 and AUC, the best values are obtained by the combinations that are created by applying both rules to the top 5 models estimated with test set 1. That is, the combination of Model 3 & 5. Moreover, as stated by Table 23 in the appendix, the best prediction performance for test set 2 is obtained via the combinations that are made by applying both rules to the 5 top models estimated by test set 2. That is the combination Model 3 & 4. The best combination for test set 3 are however obtained by applying both rules to the top 10, as stated in Table 29 in the appendix. Moreover, Table 12 shows that, on average the best combination of models for a specific test set is obtained by applying both inclusion rules to the top 5 models estimated by that specific set.

By applying the rules of inclusion, the prediction performance of the test set is used, hence make use of out-of-sample information. However, it is often the case that one cannot estimate these results, because the out-of-sample data is not available. A possible solution is making simple combinations as described above. Another solution is looking at the model combinations obtained by one test set, and evaluating these combinations with estimates from another test set. For example, the blue model combinations in Table 13 show the prediction performance of model combinations based on test set 2, estimated with test set 1. The results of the model combinations estimated with test set 2 and 3 are displayed in the appendix in Table 23 and Table 29. Table 12 shows that, on average the combinations based on another test set outperform simple combinations for the top 5 models. A third option is creating model combinations using the train set. The analysis of these combinations is described in Section 6.1.1 in the appendix. It is concluded that these train set combinations do not outperform the other combinations, which is also shown in Table 12.

4.2.3.2 Traditional Benchmark for the Additional Model Combinations

The analysis of this comparison is described in Section 6.1.2 in the appendix. Overall, it is concluded that the BMA weights only outperform the traditional weights for larger model combinations, as this results in significantly assigning larger weights to models with a higher prediction performance. Furthermore, the performance of BMA compared to the traditional weights is dependent on the ability of BVS to order the different models on prediction performance. An increase in this ability also increases the performance of the simple BMA model combinations compared to different combinations based on rules, as a combination of perfectly ordered models weighted from high to lower is the optimal combination. Furthermore, when BVS becomes more robust for different out-of-sample sets, the combinations based on different sets or the train set approach the combinations based on the same set, increasing the performance of BMA.

4.3 Comparing BVS and BMA

Table 14: Top 10 BVS models and simple BMA combinations. Indicates the absolute difference between the best value for the criteria and the result of the model or combination, summed over all three test sets. Average (5) indicates the average criteria differences for the best 5 models or their simple combinations, where average (10) shows the averages of the best 10 models or their simple combinations.

Top 10 Models	FP60	FP90	FP50-100	AUC	Z-score
Model 1	176	75	62	0.006	71.718
Model 2	169	136	53	0.005	75.758
Model 3	137	154	28	0.003	53.183
Model 4	120	216	19	0.002	58.930
Model 5	173	71	63	0.006	73.527
Model 6	179	129	53	0.005	77.237
Model 7	92	81	60	0.006	72.883
Model 8	50	1415	1214	0.131	127.301
Model 9	108	144	50	0.004	77.287
Model 10	129	155	30	0.003	56.024
Average (5)	129	109	38	0.004	55.519
Average (10)	133	258	163	0.017	74.385
Model Combinations					
Model 1 & 2	188	105	56	0.006	73.384
Model 1, 2 & 3	173	113	53	0.005	71.540
Model 1, 2, 3, & 4	172	126	48	0.004	69.812
Model 1, 2, 3, 4 & 5	170	123	49	0.004	69.011
Model 1, 2, 3, 4, 5 & 6	174	125	49	0.004	69.866
Model 1, 2, 3, 4, 5, 6 & 7	175	120	50	0.004	69.923
Model 1, 2, 3, 4, 5, 6, 7 & 8	171	201	75	0.007	67.202
Model 1, 2, 3, 4, 5, 6, 7, 8 & 9	173	200	75	0.007	67.910
Model 1, 2, 3, 4, 5, 6, 7, 8 , 9 & 10	171	200	73	0.007	67.618
Average (5)	176	117	52	0.005	70.937
Average (10)	174	146	59	0.005	69.585

Table 14 displays the differences between the optimal criteria values and the values of the selected models and combinations summed over the three test sets. It is concluded that the optimal individual BVS models outperforms the optimal BMA combination in terms of prediction performance. As the results might be biased because some test sets have larger differences on average, Table 30 in the appendix displays the average differences in percentages. However, there are no differences in which model or combination has the optimal criteria value. By analyzing the earlier described tables with individual models and model combinations, it is shown

that this result also holds for the test sets separately. Hence, it is concluded that the optimal BVS performing models outperform the optimal BMA combinations.

However, BVS orders the models on their posterior probability, which in practice does not consistently result in the models being perfectly ordered on prediction performance for every test set. For example Model 1, with the highest posterior probability, is on average outperformed by the averages of the BMA combinations. Consequently, when predicting out-of-sample, one does not know exactly which models will have higher prediction performance and which ones will not. For example as stated in Table 14, Model 10 has shows strong prediction power over the three test sets, whereas Model 8 has been assigned higher posterior probability by BVS, but has significantly lower prediction performance.

As stated in the introduction, BVS can result in two different models that are comparably likely, but have different explanatory variables, regression coefficients and consequently predictions for the PD. As both models might be over-fitting on some aspects of the training set, averaging them results in a more robust and stable estimate. This is illustrated with an example focusing on Model 1 and Model 2. The Bayes factor of model 1 over model 2 is calculated as $BF_{35} = \frac{0.328}{0.231} = 1.42$. Given that one of the two models is correct, we have $P(M_3) = \frac{1.42}{2.42} = 0.0.587$, and $P(M_5) = 0.413$, indicating that their likeliness does not significantly differ. Table 15 displays the prediction performance criteria of these two models, and their BMA combination, for all three sets separately and the average over the sets. As shown in Table 6, the regression coefficients are almost identical for the variables that are included in both models. The difference is that Model 2 also consists of the variable CLTV. Hence by weighting the models, a weighted variable CLTV is included. For test set 1, Model 1 has higher average prediction performance, for test set 2 Model 2 has higher average prediction performance, and for test set 3 Model 1 has only slightly higher average prediction performance. Consequently, the combination of the two models almost always has an average prediction performance, and for test set 3 even the best prediction performance. With the use of BMA, the BVS models are averaged, resulting in a loss in accuracy for the best models, but increased robustness and stability in the prediction performance. Hence, possible outliers in the data or in models will also have less influence on the results obtained by BMA. This is also shown in Table 14, where the differences in prediction criteria values between the different models are clearly larger than the differences between the combinations. For example, the values for FP50-100 range from 48 to 75 for the different BMA combinations, whereas they range from 19 to 1214 for the individual models.

Table 15: The results of model 1, model 2 and the BMA combination of model 1 and 2, for all three test sets and the average over the test sets.

Set 1		PosProb	FP60	FP90	FP50-100	AUC	Z-score
Model 1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.328	283	2460	1283	0.890	126.119
Model 2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.231	288	2488	1291	0.889	126.655
Model 1 & 2			287	2476	1287	0.889	126.333
Set 2							
Model 1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.328	781	3863	2072	0.822	366.512
Model 2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.231	787	3879	2057	0.824	369.550
Model 1 & 2			782	3873	2065	0.823	368.022
Set 3							
Model 1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.328	1094	1548	1507	0.857	131.863
Model 2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.231	1076	1565	1505	0.857	132.329
Model 1 & 2			1101	1552	1504	0.857	131.805
Average							
Model 1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.328	719	2624	1621	0.856	208.164
Model 2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.231	717	2644	1618	0.856	209.511
Model 1 & 2			723	2634	1619	0.856	208.720

In Section 4.2, it is stated that an increase in the ability of BVS to order the models on their prediction performance results in a better performance of BMA in contrast to the benchmark weights. This result remains valid, nevertheless this inability is the reason that BMA results in more robust and stable predictions compared to BVS. When BVS would result in one model that robustly outperforms all other models, combining this model would always result in a loss in prediction performance. However, the existence of such a model is dependent on the data set, and is for almost every set not available.

Lastly, Table 12 displays the average criteria values for the individual models and different model combinations. The results show that the combinations created by the inclusion rules on average outperform the individual models, for both the top 5 and top 10 models. This is an important result, as it shows that for the right combinations, BMA does not only result in decreased variation in the model specific errors, but also in increased prediction performance on average.

For the practical implementation of the results, it follows from this result that BMA can best be used for the modelling of the PD of mortgage portfolios. As shown before BVS outperforms the traditional variable selection algorithms FS, BE, and SS. Additionally, BMA outperforms BVS in that it is more robust and stable, which is an improves a PD model, as it might have to deal with outliers and high variation in the observations. Consequently, for combinations created with both inclusion rules, BMA also results in higher prediction performance on average.

5 Conclusions and Extensions

This paper investigates the effects of BVS and BMA on predicting the PD of a mortgage portfolio. More specifically, it answers the question whether the use of BVS and BMA results in an increase in prediction performance when modelling the PD of mortgage portfolios. As a benchmark traditional criteria and weights are evaluated. Given the top 5 models selected with BVS, the comparison between BVS and the traditional benchmark criteria AIC and BIC is not consistent for the different test sets. Hence no strong evidence in favour of BVS is discovered. However, when considering models outside this top 5, BVS does obtain better results than the traditional benchmark criteria. The only criteria for which the benchmark models strictly outperform BVS is the Z-score, indicating that these models have better prediction performance for a TPR of 100%, that is when all defaults are required to be predicted as defaults. Nevertheless, BVS does on average outperform the traditional variable selection algorithms FS, BE and SS. Furthermore, for different test sets, the BVS models also outperform the models obtained by these traditional algorithms. Hence, BVS does increase the prediction performance of a PD model compared to the benchmark.

For BMA different combinations are compared on prediction performance and are compared to traditional weighted combinations. The best performing BMA combinations are created by applying the two inclusion rules to the result of the same test set. However, as this is not possible for out-of-sample data, applying both rules to the results of another test set is a good alternative. It results in higher prediction performance compared to creating simple model combinations. The best performing combinations are small to medium sized, ranging from two to seven included models. For these best combinations, the BMA combinations do not outperform the traditional weighting benchmark. However, for larger combinations with models from different places in the top 10, BMA does outperform the benchmark.

The optimal individual BVS models outperform the optimal BMA combinations. Nevertheless, BMA combinations decrease the variation in the model specific errors, increases robustness for different test sets, and on average outperform the BVS models on prediction performance. Better ability of BVS to robustly order the models on prediction performance would result in better performance of BMA compared to the benchmark criteria, but would also decrease the overall effect of averaging models, as the optimal model can simply be selected.

Hence in short, BVS outperforms the benchmark in selecting the optimal subset of included explanatory variables, BMA does not significantly outperform the benchmark, but does on average outperform the prediction performance of BVS for model combinations created with both inclusion rules. Hence, it is concluded that applying BVS does increase the prediction performance of the PD model, and making simple combinations of the optimal models, with for example BMA, results in a PD model with improved prediction performance and less variation in the model specific errors. The use of this methodology does significantly decrease the model specific error and increases the prediction performance of models used to predict the PD of a mortgage portfolio.

Further research could investigate the wider application of the obtained result in the field of credit risk. The sample analysed in this paper is relatively small compared to the population

of mortgage loans. Additionally, data from four out of nineteen available years is used. Further research could potentially investigate larger data sets to increase the general application of the results. This can include more observations, investigating a longer time span of the loan, or analysing more years of loan origin. Furthermore, different explanatory variables could be included. For example, when a longer time span of the loan is examined, the variable indicating arrear payments in the past can be included. Moreover, when analysing multiple years of origin, a variable describing the underlying business cycle could be included.

Another possible extension is the comparison of SSVS with other BVS methods. As described in Section 1, literature describes a wide variety of different BVS methods. This paper investigates SSVS, however further research could also investigate the application of for example MCMCRJ for modelling the PD. These obtained results of the different models can be compared on prediction performance, but also on for example the speed of the method, or the ease of adapting it to a different set-up or topic.

Furthermore, this paper analyses traditional variable selection methods and criteria like FS or AIC as a benchmark. Existing literature also describes a variety of different new methods that can be used as a benchmark. For example, machine learning techniques can be used to create PD models. Further research can investigate the differences of BVS and these methods. This can be done on efficiency or predictive performance, but also applicability due to differences in for example transparency and subjectivity.

References

Abdi, H. (2007). Binomial distribution: Binomial and sign tests. *Encyclopedia of measurement and statistics*, *1*.

Aguais, S. D., Forest Jr, L. R., Wong, E. Y., & Diaz-Ledezma, D. (2004). Point-in-time versus through-the-cycle ratings. *The Basel handbook: a guide for financial practitioners*.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, *23*(4), 589–609.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of accounting research*, 71–111.

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. *Critical care*, *9*(1), 112.

Casella, G., & George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, *46*(3), 167–174.

Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: Theory and implementation. *Available at czep. net/stat/mlelr. pdf*.

Fragoso, T. M., Bertoli, W., & Louzada, F. (2018). Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*, *86*(1), 1–28.

George, E. I., & McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society: Series B (Methodological)*, *46*(2), 149–170.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, *82*(4), 711–732.

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Kotz, D. M. (2009). The financial and economic crisis of 2008: A systemic crisis of neoliberal capitalism. *Review of radical political economics*, *41*(3), 305–317.

Lingo, M., & Winkler, G. (2008). Discriminatory power-an obsolete validation criterion? *Available at SSRN 1026242*.

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, *89*(428), 1535–1546.

Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106). Sage.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of bayesian variable selection methods: what, how and which. *Bayesian analysis*, *4*(1), 85–117.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109–131.

Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, *443*(7), 59–72.

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology*, *53*(5), 793–808.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, *25*, 111–164.

Raftery, A. E. (1999). Bayes factors and bic: Comment on "a critique of the bayesian information criterion for model selection". *Sociological Methods & Research*, *27*(3), 411–427.

Raftery, A. E., & Lewis, S. (1991). *How many iterations in the gibbs sampler?* (Tech. Rep.). WASHINGTON UNIV SEATTLE DEPT OF STATISTICS.

Roberts, G. O., & Smith, A. F. (1994). Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. *Stochastic processes and their applications*, *49*(2), 207–216.

Zhang, Z., Li, J., Zhao, X.-Q., Wang, J., Wong, G. K.-S., & Yu, J. (2006). Calculating ka and ks through model selection and model averaging. *Genomics, proteomics & bioinformatics*, *4*(4), 259–263.

6 Appendix

6.1 Additional results

6.1.1 Creating Combinations using the Train set

Here, the possibility of estimating the model combinations on estimates obtained by the the train data set is investigated. Table 18 in the appendix states the prediction performance obtained when estimating the top 10 models with the train set. Here it is shown that when both rules are applied to the top 5 and top 10, the combinations Model 1, 2 & 3, and Model 7 & 9 are obtained respectively. These are the same models obtained by test set 3, and are presented

in Table 29. Consequently, for estimates obtained with test set 3, the prediction performance is high. However, for the other two test sets the estimates show lower prediction performance. As displayed in Table 12, on average the prediction performance of the combinations obtained by the train set do not outperform the other above described combinations.

6.1.2 Analysis of the Traditional Benchmark Criteria for the Additional Model Combinations

Table 16: Prediction performance of different AIC weighted combinations estimated with test set 1. The different combinations are based on applying the model inclusion rules to the top 10 models estimated with test set 1, 2 3.

AIC Combinations	FP60	FP90	FP50-100	AUC	Z-score
Model 1 & 2	287	2478	1288	0.889	126.494
Model 1, 2 & 3	281	2479	1283	0.890	126.226
All models	261	2595	1299	0.888	124.329
Combinations based on set 1					
Model 1, 3, 4 & 5	281	2479	1280	0.890	126.280
Model 1, 3, 4, 5 & 10	280	2478	1278	0.89	126.280
Model 1, 3 & 5	281	2465	1279	0.890	126.066
Model 1, 3, 5, 7, 8, 9 & 10	255	2581	1301	0.888	123.145
Model 3 & 5	280	2467	1276	0.890	126.119
Model 1, 3, 5 & 10	280	2467	1276	0.890	126.119
Combinations based on set 2					
Model 2, 3 & 4	281	2494	1282	0.890	126.602
Model 2, 3, 4, 6, 9 & 10	281	2488	1283	0.890	126.871
Model 1, 2, 3 & 4	281	2489	1283	0.890	126.387
Model 1, 2, 3, 4, 6, 7, 8 & 9	260	2598	1303	0.888	123.760
Model 3 & 4	280	2492	1279	0.890	126.333
Model 2, 3, 4, 6 & 9	281	2491	1285	0.890	126.925
Combinations based on set 3					
Model 2, 3 & 4	281	2494	1282	0.890	126.602
Model 2, 4, 7 & 9	285	2488	1288	0.889	126.979
Model 1, 2, 3 & 4	281	2489	1283	0.890	126.387
Model 7, 8 & 9	282	2573	1339	0.884	119.473
Model 3 & 4	280	2492	1279	0.890	126.333
Model 7 & 9	290	2480	1292	0.889	127.196

For the other two test sets, results of the model combinations weighted by AIC, BIC and equal weights are again only marginal different from each other, where the AIC weights show marginally better results. The results of the AIC weights estimated with test set 1, for the different combinations created by all three test sets, are located in Tabel 16. Here it is shown that the average prediction performance of the best benchmark combinations is similar to the best BMA combinations, displayed in Table 10. This is a general result, as it is confirmed by the results of the same combinations estimated for the other two sets as shown in the appendix. An explanation for this is that most best performing combinations are made out of models that are ordered closely to each other, resulting in small differences in weights and hence results that are comparable to the traditional weights. If we focus on the model combinations with a large number of models, the BMA models outperform the traditional weights, e.g., the combinations

'All Models', and Model 1, 3, 5, 7, 8, 9 & 10. Overall, it is concluded that the BMA weights only outperform the traditional weights for larger model combinations, as this results in significantly assigning larger weights to models with a higher prediction performance.

A general explanation why BMA does not outperform the traditional weights in general, as stated in Section 4.1, is that BVS is able to select top performing models, but does not perfectly order them on prediction performance. As stated earlier, combining individual models with high prediction performance results in combinations with high prediction performance. Furthermore, combinations of models with low prediction performance and high posterior probabilities with models with high prediction performance and low posterior probabilities is outperformed by an equally weighted combination of these models. Hence, the effectiveness of BVS has a direct influence on the effectiveness of BMA. Another factor that should be kept in mind is that we evaluate out-of-sample prediction performance. That is, the results are in some way always dependent on the comparability of the train set and the test set. A clear example is the simple combination Model 1 & 2. For test set 1, Model 1 has lower prediction performance, resulting in the traditional weights outperforming BMA as BMA assigns a higher weight to Model 1. However, for test set 3, Model 1 has higher prediction performance resulting in BMA outperforming the traditional weights. This would not be possible if prediction the performance of the models was consistent for different test sets. This is however partially dependent on the effectiveness of BVS, and partially on the specific data set.

6.2 Train Set

Table 17: Comparison of the regression coefficients of the top 5 models estimate with BVS and IRLS. 'Bayesian' states the regression coefficients estimated using BVS, 'Traditional' states the regression coefficients estimated with IRLS. 'SE' are the IRLS standard errors in parenthesis. The Bayesian estimates do not differ significantly from the IRLS estimates, for a 95% (traditional) significance interval.

	Statistic	fico	mi_pct	cnt_units	cltv	dti	orig_upb	ltv	int_rt	cnt_borr
Model 1	Bayesian	-0.012	0.062	-0.408					1.378	-1.07
	Traditional	-0.010	0.054	-0.160					1.509	-1.115
	SE	(0.003)	(0.014)	(1.044)					(0.288)	(0.443)
Model 2	Bayesian	-0.013	0.061	-0.319	-0.047				1.385	-1.024
	Traditional	-0.010	0.054	-0.156	0.008				1.510	-1.116
	SE	(0.003)	(0.014)	(1.045)	(0.058)				(0.289)	(0.443)
Model 3	Bayesian	-0.014	0.076	-0.315				-0.014	1.324	-1.033
	Traditional	-0.010	0.051	-0.155				0.003	1.508	-1.115
	SE	(0.003)	(0.024)	(1.045)				(0.024)	(0.288)	(0.443)
Model 4	Bayesian	-0.014	0.079	-0.223	-0.045			-0.017	1.264	-0.942
	Traditional	-0.010	0.052	-0.151	0.008			0.003	1.508	-1.115
	SE	(0.003)	(0.024)	(1.046)	(0.059)			(0.024)	(0.289)	(0.443)
Model 5	Bayesian	-0.013	0.063	-0.331		-0.009			1.346	-1.031
	Traditional	-0.010	0.054	-0.162		0.001			1.507	-1.114
	SE	(0.003)	(0.014)	(1.045)		(0.016)			(0.290)	(0.444)

Table 18: Prediction performance of the top 10 models with the highest posterior probability, as calculated with BVS for test the train set.

	Variables	PosProb	AIC	BIC	FP60	FP90	FP50-100	AUC	Z-score
Model 1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.328	291.170	331.367	279	1562	1234	0.896	165.63
Model 2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.231	293.152	340.049	281	1500	1233	0.896	166.68
Model 3	Credit Score, Insurance, Units, LTV, Interest Rate, Borrowers	0.118	293.154	340.050	275	1522	1235	0.896	176.21
Model 4	Credit Score, Insurance, Units, CLTV, LTV, Interest Rate, Borrowers	0.077	295.136	348.732	276	1468	1234	0.896	177.09
Model 5	Credit Score, Insurance, Units, DTI, Interest Rate, Borrowers	0.077	293.167	340.064	276	1568	1237	0.896	166.20
Model 6	Credit Score, Insurance, Units, CLTV, DTI, Interest Rate, Borrowers	0.054	295.149	348.746	276	1505	1234	0.896	166.88
Model 7	Credit Score, Insurance, Interest Rate, Borrowers	0.048	289.194	322.692	281	1420	1226	0.896	165.35
Model 8	Insurance, Units, Interest Rate, Borrowers	0.026	300.384	333.882	337	1828	1361	0.886	371.10
Model 9	Credit Score, Insurance, CLTV, Interest Rate, Borrowers	0.025	291.175	331.372	282	1354	1224	0.896	166.20
Model 10	Credit Score, Insurance, Units, DTI, LTV, Interest Rate, Borrowers	0.016	295.151	348.747	271	1532	1237	0.896	176.87

6.3 Results Test Set 1

6.3.1 Bayesian Variable Selection

Table 19: results for $c = 5M$, with the same models and prediction performance but different posterior probability as $c = 500K$.

	M_l	Top 5 models	—PosProb	FN60	FN90	FN50-100	AUC	Z-value
$c = 50M$	M_1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.415	283	2460	1283	0.890	126.119
	M_2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.307	288	2488	1291	0.889	126.655
	M_3	Credit Score, Insurance, Units, LTV, Interest Rate, Borrowers	0.122	280	2479	1274	0.890	125.959
	M_4	Credit Score, Insurance, Units, CLTV, LTV, Interest Rate, Borrowers	0.080	280	2503	1282	0.890	126.871
	M_5	Credit Score, Insurance, Units, DTI, Interest Rate, Borrowers	0.076	282	2454	1279	0.890	126.119

6.4 Test set 2

6.4.1 Bayesian Variable Selection

Table 20: Top 5 models with the highest posterior probability given $p = 0.90$ and $c = 500,000$, and their corresponding prediction performance criteria. The included explantory variables, AIC and BIC are calculated by the training set. The prediction performance criteria are calculated with test set 2. PosProb = Posterior probability, FP60 = amount of FP observations given a TRP rate of 60%, FP90 = amount of FP observations given a TRP rate of 90%, FP50-100 = average amount of FP observations in the TPR range of 50% – 100%, Z-value based on a TPR of 100%.

	Model	AIC	BIC	FN60	FN90	FN50-100	AUC	Z-value
BVS model 1	Credit Score, Insurance, Units, Interest Rate, Borrowers	291.170	331.367	781	3863	2072	0.822	366.512
BVS model 2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	293.152	340.049	787	3879	2057	0.824	369.550
BVS model 3	Credit Score, Insurance, Units, LTV, Interest Rate, Borrowers	293.154	340.050	761	3799	2048	0.825	350.978
BVS model 4	Credit Score, Insurance, Units, CLTV, LTV, Interest Rate, Borrowers	295.136	348.732	770	3820	2033	0.826	355.029
BVS model 5	Credit Score, Insurance, Units, DTI, Interest Rate, Borrowers	293.167	340.064	779	3870	2075	0.822	368.784
FS model	Credit Score, Insurance, Units, CLTV, DTI, UPB, LTV, Interest Rate, Borrowers	287.398	354.393	539	4767	2447	0.791	186.556
BE & SS model	Credit Score, Insurance, UPB, Interest Rate, Borrowers	280.327	320.524	502	4795	2430	0.792	205.176

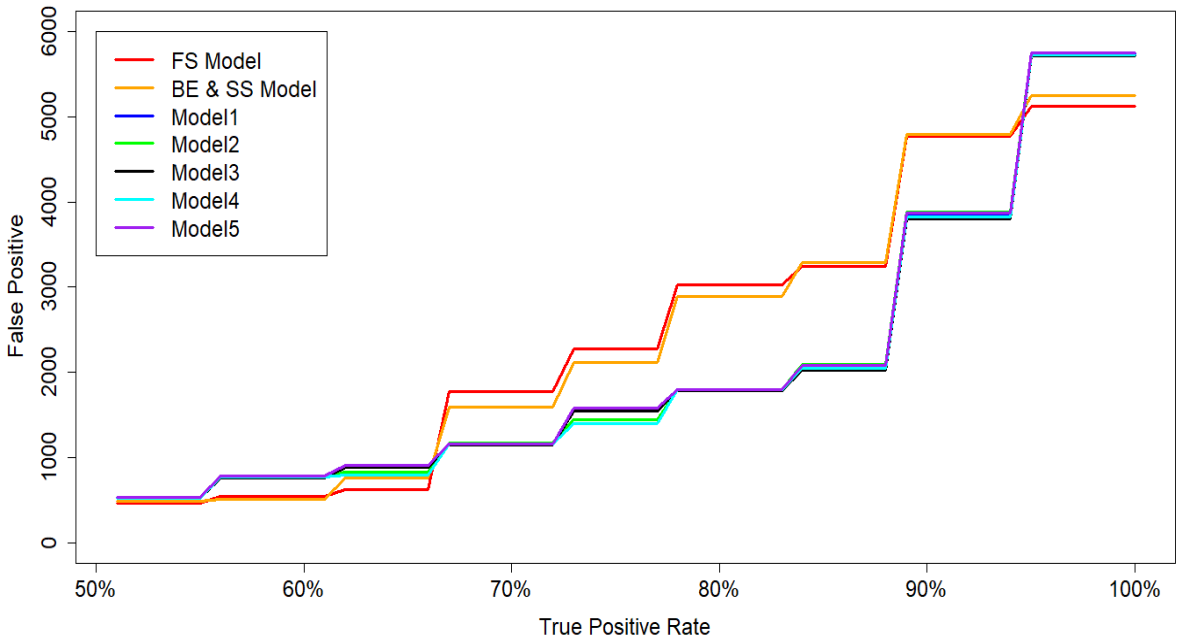


Figure 7: Amount of FP given the TPR plotted in the range of 50% – 100%. Based on test set 2.

Table 21: Prediction performance of the top 10 models with the highest posterior probability, as calculated with BVS. Estimated with test set 2.

	Variables	PosProb	AIC	BIC	FP60	FP90	FP50-100	AUC	Z-score
Model 1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.328	291.170	331.367	781	3863	2072	0.822	366.512
Model 2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.231	293.152	340.049	787	3879	2057	0.824	369.550
Model 3	Credit Score, Insurance, Units, LTV, Interest Rate, Borrowers	0.118	293.154	340.050	761	3799	2048	0.825	350.978
Model 4	Credit Score, Insurance, Units, CLTV, LTV, Interest Rate, Borrowers	0.077	295.136	348.732	770	3820	2033	0.826	355.029
Model 5	Credit Score, Insurance, Units, DTI, Interest Rate, Borrowers	0.077	293.167	340.064	779	3870	2075	0.822	368.784
Model 6	Credit Score, Insurance, Units, CLTV, DTI, Interest Rate, Borrowers	0.054	295.149	348.746	785	3884	2061	0.824	371.096
Model 7	Credit Score, Insurance, Interest Rate, Borrowers	0.048	289.194	322.692	786	3866	2074	0.822	367.265
Model 8	Insurance, Units, Interest Rate, Borrowers	0.026	300.384	333.882	719	4314	2579	0.783	306.068
Model 9	Credit Score, Insurance, CLTV, Interest Rate, Borrowers	0.025	291.175	331.372	791	3885	2059	0.824	370.321
Model 10	Credit Score, Insurance, Units, DTI, LTV, Interest Rate, Borrowers	0.016	295.151	348.747	754	3807	2051	0.824	353.664

6.4.2 Bayesian Model Averaging

Table 22: Prediction performance of analysed different simple combinations of models, estimated using test set 2, weighted by BMA.

BMA simple combinations	FP60	FP90	FP50-100	AUC	Z-score
Model 1 & 2	782	3873	2065	0.823	368.022
Model 1, 2 & 3	777	3855	2062	0.823	366.512
Model 1, 2, 3 & 4	777	3853	2058	0.824	365.019
Model 1, 2, 3, 4 & 5	776	3853	2059	0.824	364.279
Model 1, 2, 3, 4, 5 & 6	777	3858	2059	0.824	365.019
Model 1, 2, 3, 4, 5, 6 & 7	777	3857	2060	0.824	365.019
Model 1, 2, 3, 4, 5, 6, 7 & 8	784	3884	2071	0.823	359.926
Model 1, 2, 3, 4, 5, 6, 7, 8 & 9	785	3883	2071	0.823	360.641
Model 1, 2, 3, 4, 5, 6, 7, 8 , 9 & 10	783	3882	2070	0.823	360.641

Table 23: Prediction performance of different BMA combinations estimated with test set 2. The different combinations are based on applying the model inclusion rules to the top 10 models estimated with test set 1, 2 3.

BMA combinations	FP60	FP90	FP50-100	AUC	Z-score
Model 1 & 2	782	3873	2065	0.823	368.022
Model 1, 2 & 3	777	3855	2062	0.823	366.512
All models	783	3882	2070	0.823	360.641
Combinations based on set 1					
Model 1, 3, 4 & 5	776	3845	2062	0.823	363.543
Model 1, 3, 4, 5 & 10	775	3843	2061	0.824	362.084
Model 1, 3 & 5	775	3846	2066	0.823	362.811
Model 1, 3, 5, 7, 8, 9 & 10	781	3891	2080	0.822	357.804
Model 3 & 5	766	3825	2058	0.824	357.104
Model 1, 3, 5 & 10	775	3846	2067	0.823	362.811
Combinations based on set 2					
Model 2, 3 & 4	776	3850	2049	0.825	363.543
Model 2, 3, 4, 6, 9 & 10	778	3852	2050	0.825	363.543
Model 1, 2, 3 & 4	777	3853	2058	0.824	365.019
Model 1, 2, 3, 4, 6, 7, 8 & 9	784	3886	2070	0.823	360.641
Model 3 & 4	764	3808	2039	0.826	352.987
Model 2, 3, 4, 6 & 9	778	3855	2051	0.824	364.279
Combinations based on set 3					
Model 2, 3 & 4	776	3850	2049	0.825	363.543
Model 2, 4, 7 & 9	781	3861	2055	0.824	368.022
Model 1, 2, 3 & 4	777	3853	2058	0.824	365.019
Model 7, 8 & 9	821	4059	2167	0.815	344.506
Model 3 & 4	764	3808	2039	0.826	352.987
Model 7 & 9	786	3876	2068	0.823	368.784

Table 24: Prediction performance of different AIC weighted combinations estimated with test set 2. The different combinations are based on applying the model inclusion rules to the top 10 models estimated with test set 1,2 3.

AIC combinations	FP60	FP90	FP50-100	AUC	Z-score
Model 1 & 2	783	3874	2063	0.823	368.784
Model 1, 2 & 3	775	3846	2057	0.824	363.543
All models	795	3940	2096	0.821	355.029
Combinations based on set 1					
Model 1, 3, 4 & 5	769	3839	2055	0.824	359.215
Model 1, 3, 4, 5 & 10	766	3832	2054	0.824	358.507
Model 1, 3 & 5	772	3838	2064	0.823	361.361
Model 1, 3, 5, 7, 8, 9 & 10	800	3973	2116	0.819	349.656
Model 3 & 5	767	3830	2060	0.824	357.804
Model 1, 3, 5 & 10	767	3830	2060	0.824	357.804
Combinations based on set 2					
Model 2, 3 & 4	774	3835	2045	0.825	358.507
Model 2, 3, 4, 6, 9 & 10	774	3848	2050	0.825	359.926
Model 1, 2, 3 & 4	772	3844	2050	0.824	361.361
Model 1, 2, 3, 4, 6, 7, 8 & 9	803	3967	2103	0.82	352.314
Model 3 & 4	765	3811	2038	0.826	353.664
Model 2, 3, 4, 6 & 9	778	3853	2051	0.824	363.543
Combinations based on set 3					
Model 2, 3 & 4	774	3835	2045	0.825	358.507
Model 2, 4, 7 & 9	781	3861	2055	0.824	367.264
Model 1, 2, 3 & 4	772	3844	2050	0.824	361.361
Model 7, 8 & 9	797	4117	2198	0.812	334.814
Model 3 & 4	765	3811	2038	0.826	353.664
Model 7 & 9	787	3880	2065	0.823	369.55

6.5 Test set 3

6.5.1 Bayesian Variable Selection

Table 25: Top 5 models with the highest posterior probability given $p = 0.90$ and $c = 500,000$, and their corresponding prediction performance criteria. The included explantory variables, AIC and BIC are calculated by the training set. The prediction performance criteria are calculated with test set 3. PosProb = Posterior probability, FP60 = amount of FP observations given a TRP rate of 60%, FP90 = amount of FP observations given a TRP rate of 90%, FP50-100 = average amount of FP observations in the TPR range of 50% − 100%, Z-value based on a TPR of 100%.

	Model	AIC	BIC	FN60	FN90	FN50-100	AUC	Z-value
BVS model 1	Credit Score, Insurance, Units, Interest Rate, Borrowers	291.170	331.367	1094	1548	1507	0.857	131.863
BVS model 2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	293.152	340.049	1076	1565	1505	0.857	132.329
BVS model 3	Credit Score, Insurance, Units, LTV, Interest Rate, Borrowers	293.154	340.050	1078	1672	1506	0.857	129.022
BVS model 4	Credit Score, Insurance, Units, CLTV, LTV, Interest Rate, Borrowers	295.136	348.732	1052	1689	1504	0.857	129.806
BVS model 5	Credit Score, Insurance, Units, DTI, Interest Rate, Borrowers	293.167	340.064	1094	1543	1509	0.857	131.400
FS model	Credit Score, Insurance, Units, CLTV, DTI, UPB, LTV, Interest Rate, Borrowers	287.398	354.393	878	2784	1709	0.842	74.677
BE & SS model	Credit Score, Insurance, UPB, Interest Rate, Borrowers	280.327	320.524	771	2637	1617	0.850	75.130

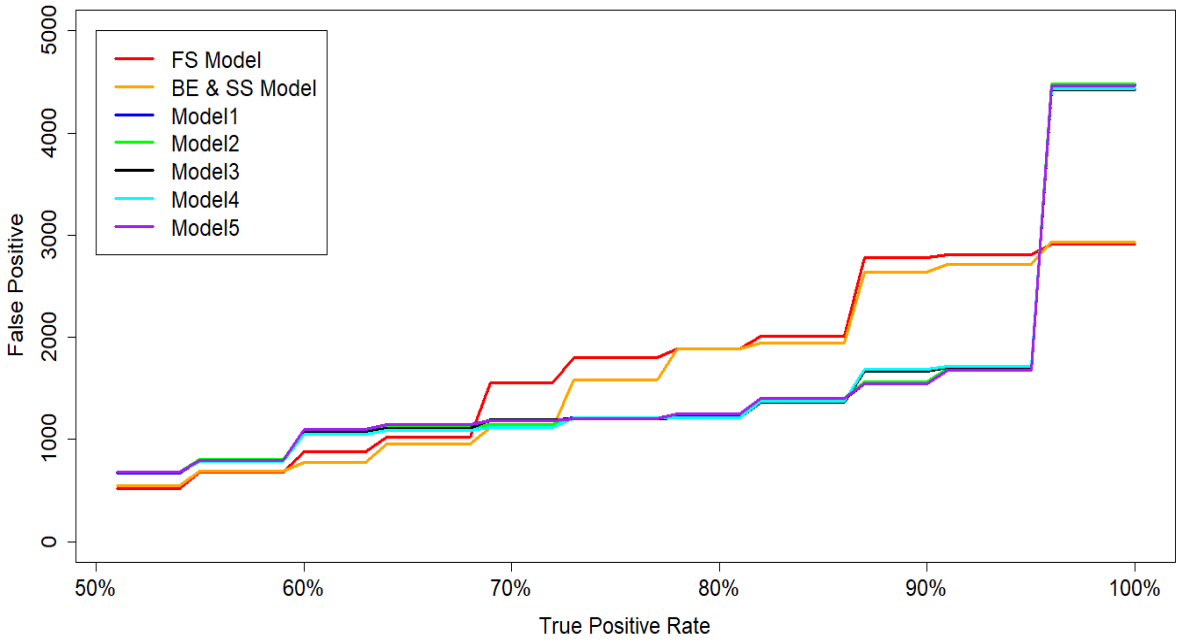


Figure 8: Indicates the amount of FP given the TPR plotted in the range of 50% – 100%. Based on test set 3.

Table 26: Prediction performance of the top 10 models with the highest posterior probability, as calculated with BVS. Estimated with test set 3.

	Variables	PosProb	AIC	BIC	FP60	FP90	FP50-100	AUC	Z-score
Model 1	Credit Score, Insurance, Units, Interest Rate, Borrowers	0.328	291.170	331.367	1094	1548	1507	0.857	131.863
Model 2	Credit Score, Insurance, Units, CLTV, Interest Rate, Borrowers	0.231	293.152	340.049	1076	1565	1505	0.857	132.329
Model 3	Credit Score, Insurance, Units, LTV, Interest Rate, Borrowers	0.118	293.154	340.050	1078	1672	1506	0.857	129.022
Model 4	Credit Score, Insurance, Units, CLTV, LTV, Interest Rate, Borrowers	0.077	295.136	348.732	1052	1689	1504	0.857	129.806
Model 5	Credit Score, Insurance, Units, DTI, Interest Rate, Borrowers	0.077	293.167	340.064	1094	1543	1509	0.857	131.400
Model 6	Credit Score, Insurance, Units, CLTV, DTI, Interest Rate, Borrowers	0.054	295.149	348.746	1094	1559	1506	0.857	132.154
Model 7	Credit Score, Insurance, Interest Rate, Borrowers	0.048	289.194	322.692	999	1548	1499	0.858	131.631
Model 8	Insurance, Units, Interest Rate, Borrowers	0.026	300.384	333.882	985	3306	2146	0.798	256.157
Model 9	Credit Score, Insurance, CLTV, Interest Rate, Borrowers	0.025	291.175	331.372	1009	1565	1496	0.858	132.329
Model 10	Credit Score, Insurance, Units, DTI, LTV, Interest Rate, Borrowers	0.016	295.151	348.747	1079	1667	1508	0.857	128.856

6.5.2 Bayesian Model Averaging

Table 27: Prediction performance of analysed different simple combinations of models, estimated using test set 3, weighted by BMA.

BMA simple combinations	FP60	FP90	FP50-100	AUC	Z-score
Model 1 & 2	1101	1552	1504	0.857	131.805
Model 1, 2 & 3	1096	1575	1506	0.857	131.631
Model 1, 2, 3 & 4	1096	1589	1506	0.857	131.343
Model 1, 2, 3, 4 & 5	1095	1586	1506	0.857	131.228
Model 1, 2, 3, 4, 5 & 6	1098	1583	1506	0.857	131.343
Model 1, 2, 3, 4, 5, 6 & 7	1099	1580	1506	0.857	131.400
Model 1, 2, 3, 4, 5, 6, 7 & 8	1090	1612	1516	0.856	133.986
Model 1, 2, 3, 4, 5, 6, 7, 8 & 9	1091	1612	1516	0.856	133.926
Model 1, 2, 3, 4, 5, 6, 7, 8 , 9 & 10	1091	1614	1516	0.856	133.687

Table 28: Prediction performance of different BMA combinations estimated with test set 3. The different combinations are based on applying the model inclusion rules to the top 10 models estimated with test set 1,2 3.

BMA combinations	FP60	FP90	FP50-100	AUC	Z-score
Model 1 & 2	1101	1552	1504	0.857	131.805
Model 1, 2 & 3	1096	1575	1506	0.857	131.631
All models	1091	1614	1516	0.856	133.687
Combinations based on set 1					
Model 1, 3, 4 & 5	1090	1590	1507	0.857	130.712
Model 1, 3, 4, 5 & 10	1090	1594	1507	0.857	130.655
Model 1, 3 & 5	1091	1578	1509	0.857	131.113
Model 1, 3, 5, 7, 8, 9 & 10	1086	1625	1524	0.855	134.648
Model 3 & 5	1085	1620	1507	0.857	130.144
Model 1, 3, 5 & 10	1091	1582	1509	0.857	131.113
Combinations based on set 2					
Model 2, 3 & 4	1097	1617	1505	0.857	131.17
Model 2, 3, 4, 6, 9 & 10	1097	1613	1504	0.857	131.113
Model 1, 2, 3 & 4	1096	1589	1506	0.857	131.343
Model 1, 2, 3, 4, 6, 7, 8 & 9	1090	1616	1515	0.856	134.046
Model 3 & 4	1082	1680	1505	0.857	129.301
Model 2, 3, 4, 6 & 9	1097	1610	1504	0.857	131.113
Combinations based on set 3					
Model 2, 3 & 4	1097	1617	1505	0.857	131.17
Model 2, 4, 7 & 9	1087	1588	1503	0.857	132.037
Model 1, 2, 3 & 4	1096	1589	1506	0.857	131.343
Model 7, 8 & 9	1018	1904	1604	0.847	149.465
Model 3 & 4	1082	1680	1505	0.857	129.301
Model 7 & 9	1003	1551	1497	0.858	131.805

Table 29: Prediction performance of different AIC weighted combinations estimated with test set 3. The different combinations are based on applying the model inclusion rules to the top 10 models estimated with test set 1,2 3.

AIC Combinations	FP60	FP90	FP50-100	AUC	Z-score
Model 1 & 2	1102	1555	1505	0.857	131.921
Model 1, 2 & 3	1091	1596	1505	0.857	130.941
All models	1075	1732	1547	0.853	138.572
Combinations based on set 1					
Model 1, 3, 4 & 5	1089	1617	1507	0.857	130.428
Model 1, 3, 4, 5 & 10	1087	1624	1507	0.857	130.201
Model 1, 3 & 5	1088	1591	1508	0.857	130.655
Model 1, 3, 5, 7, 8, 9 & 10	1062	1789	1566	0.851	141.32
Model 3 & 5	1086	1608	1509	0.857	130.314
Model 1, 3, 5 & 10	1086	1608	1509	0.857	130.314
Combinations based on set 2					
Model 2, 3 & 4	1094	1641	1504	0.857	130.371
Model 2, 3, 4, 6, 9 & 10	1098	1621	1504	0.857	130.655
Model 1, 2, 3 & 4	1095	1620	1505	0.857	130.541
Model 1, 2, 3, 4, 6, 7, 8 & 9	1074	1760	1554	0.852	140.524
Model 3 & 4	1083	1682	1504	0.857	129.357
Model 2, 3, 4, 6 & 9	1098	1613	1503	0.857	131.228
Combinations based on set 3					
Model 2, 3 & 4	1094	1641	1504	0.857	130.371
Model 2, 4, 7 & 9	1060	1593	1501	0.858	131.631
Model 1, 2, 3 & 4	1095	1620	1505	0.857	130.541
Model 7, 8 & 9	977	2001	1642	0.843	155.019
Model 3 & 4	1083	1682	1504	0.857	129.357
Model 7 & 9	1003	1556	1497	0.858	131.921

6.6 Results over all three test sets

Table 30: Top 10 BVS models and simple BMA combinations. Indicates the average percentage difference between the best value for the criteria and the result of the model or combination, summed over all three test sets. Average (5) indicates the average criteria differences for the best 5 models or their simple combinations, where average (10) shows the averages of the best 10 models or their simple combinations.

Top 10 Models (in %)	FP60	FP90	FP50-100	AUC	Z-score
Model 1	7.20	0.80	1.20	0.20	9.70
Model 2	7.40	1.60	1.10	0.20	10.30
Model 3	5.30	3.10	0.50	0.10	7.20
Model 4	4.90	4.00	0.50	0.10	8.10
Model 5	6.90	0.60	1.20	0.20	9.80
Model 6	7.20	1.50	1.10	0.20	10.50
Model 7	4.90	0.80	1.20	0.20	9.90
Model 8	6.00	44.50	31.80	5.10	32.90
Model 9	5.60	1.70	1.10	0.20	10.60
Model 10	4.80	3.10	0.60	0.10	7.60
Average (5)	6.34	2.02	0.90	0.16	9.02
Average (10)	6.02	6.17	4.03	0.66	11.66
Model Combinations (in %)					
Model 1 & 2	7.90	1.10	1.10	0.20	9.90
Model 1, 2 & 3	6.90	1.50	1.10	0.20	9.70
Model 2, 3, & 4	6.80	1.80	1.00	0.20	9.40
Model 1, 2, 3, 4 & 5	6.70	1.80	1.00	0.20	9.30
Model 1, 2, 3, 4, 5 & 6	6.90	1.70	1.00	0.20	9.40
Model 1, 2, 3, 4, 5, 6 & 7	6.90	1.60	1.00	0.20	9.50
Model 1, 2, 3, 4, 5, 6, 7 & 8	6.70	2.90	1.50	0.30	9.50
Model 1, 2, 3, 4, 5, 6, 7, 8 & 9	6.80	2.90	1.50	0.30	9.60
Model 1, 2, 3, 4, 5, 6, 7, 8 , 9 & 10	6.70	2.90	1.50	0.30	9.50
Average (5)	7.08	1.55	1.05	0.20	9.58
Average (10)	6.92	2.02	1.19	0.23	9.53