Bayesian updating of generic scoring models

Arkadiusz Ziemba
StatConsulting Ltd., Warsaw, Poland
www.statconsulting.biz

1. Introduction

Rationale for the method

Scoring models are among commonly used analytical tools for predicting customer's future behavior. Usually, empirical scoring models specific to a product or situation are prefered over generic scoring models. However, problems with data quality and quantity may make the construction of data-based models difficult if not impossible. In such a case modification of an existing generic scoring model becomes a viable option. In the following examples taken from the banking industry updating an existing scoring model may be the prefered solution:

- An international bank opening a branch in a new country. The bank already has a scoring model based on empirical data. Even if this model works well in the country of origin, it needs changes that will adapt it to a new economic environment.
- A local bank with a small number of customers. The bank may want to include information specific for its area of activity. This can be implemented by updating a generic scoring card using information on bank's customers.
- A financial institution offering new services and a scoring model for the services it is already offering. The institution may want to modify its model in order.
- Banks offering services to a new group of customers.

This paper presents several methods of including new information into a *generic scoring model*. In the text the term *generic scoring model* is used to denote an existing model which is regarded as the main source of prior information. The described methods may be useful in any situation where knowledge embedded in an old scoring model should not be ignored and it is advisable to create a new scoring model by modification of the old one.

Institutions already using scoring models may benefit from the proposed methods by combining different sources of information including available expert knowledge to create improved scoring models. The described approach can also be used to automatically update the model once new information becomes available.

Important advantage of using modification of a priori model could be found in consistent presentation of subsequent model. This approach could be used as effective tool against flat-maximum effect encountered in the area of statistical modelling. The flat-maximum effect refers to the situation when it is possible to create many different models to describe the same phenomenon with the similar predictive power.

It is easy to explain how flat-maximum effect works in case of classification trees. Suppose we have two predictive variables as the main potential predictors in the model with almost the same predictive power. It means those variables are challenging to split observations in the root of the tree. Slight modification of data set or criterion of the split could cause switching from one variable to another. Changing splitting variable in the root of the tree could produce totally different model by small change of the data set. More formally

we would say that model's structure is highly sensitive to changes of the data. In practise it means that updating a scoring model that is based on data collected over two years by information about new customer gained within one month could produce totally different estimations of model parameters. It could also lead to the usage of different predictive variables and such situation could cause some implementation problems. Introducing new predictive variables for application scoring enforces the necessity to change the variables supply for the scoring engine. The lack of continuity in the model representation could make it difficult to monitor changes in the business environment that are reflected in the model. In other words, we would say that descriptive role of the model could be confined by the lack of continuity.

Method overview

The article presents several methods of updating scoring models by including available prior information. We discuss updating of the most popular types of models used for credit scoring, namely linear regressions, classification trees and logistic regression models. Hence traditional scoring tables can be represented by logistic regressions, we also cover that case. The presented methods use a Bayesian framework to combine new information with prior information embedded in a generic scoring model to obtain an improved scoring model.

The idea of using a Bayesian approach to combine different sources of information is not novel. Recent examples in the credit scoring literature include Lucas (1997) and Konstantinos (2003) who used conjugate distributions (discussed later) to incorporate prior knowledge in the estimation process.

The remainder of the article is organized as follows. The next section presents the basics of Bayesian inference. Section 3 discusses simulation and data augmentation methods of including prior knowledge into a linear regression model. Section 4 describes two ways of updating classification trees: (1) re-estimating the default probability in the tree nodes and (2) modifying the tree structure. A simulation based method for updating logistic regression models is discussed in section 5. Section 6 ilustrated the working of the described methods using real data from the banking sector (results of all experiments have been obtained using the Górnik System). Finally, section 7 concludes.

2. Theoretical background of Bayesian reasoning

Classical statistics inference is directed towards using information arising from statistical experiment. So-called sample information is the only source of reliable knowledge to influence the statistical estimation of parameters we are searching for. The Bayesian approach makes an attempt to combine sample data with other relevant sources of information to improve statistical reasoning. In Bayesian statistics the other source of information beside sample information used for inference is called a **priori information**. The book of L.J. Savage (1954) covers thoroughly this subject. To use prior information operationally we need to represent it in the form of prior distribution.

The prior and sample distributions are combined by the Bayes formula:

$$p(\theta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta) * p(\theta) \qquad (1)$$

where θ is the parameter of interest, $p(\theta|\boldsymbol{y})$ denotes the conditional density of θ for given $\boldsymbol{y}$, $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a data vector of n observations, $p(\boldsymbol{y}|\theta)$ is the conditional density of $\boldsymbol{y}$ given θ and $p(\theta)$ represents the prior density of θ. The resulting distribution $p(\theta|\boldsymbol{y})$ is called a **posterior probability distribution**. We refer to $p(\boldsymbol{y}|\theta)$ as to likelihood. The aim of Bayesian analysis is to find the posterior distribution of parameter θ. If no prior information is available

we may assume a neutral prior distribution that do not favourite any values of θ. We call it a noninformative prior distribution (refer to De Groot 1970 for details).

Conjugate distributions

For a given likelihood function, a family of prior distributions is called conjugate if the posterior distribution also belongs to the same family (Diaconis, Ylvisaker, 1985)**.** The use of conjugate prior distributions offers some important practical advantages. First, they are very convenient computationally, as the evaluation of posterior distibution does not require numerical integration. Secondly, for conjugate priors it is possible to present prior information as additional data points, ie. the estimation sample is extended with additional observations that represent availible prior information. Unfortunately, conjugate priors exists only when the likelihood function belongs to the family of exponential distributions. For many practical problems (including logistic regression and hierarchical models) conjugate priors do not exists and the user has to resort to numerical integration to evaluate the posterior distribution.

Representation of prior knowledge

Let us assume that prior information is represented in a form of predictive model *(*β*, **S**)*, where β is a vector of model parameters and ***S*** stands for model structure. ***S*** could be represented as functional form of the equation describing a model if it is a parametric one or as the tree structure in case of classification trees. The form of that representation could be:

(1) enforced by the information we possess. We (potentially) know the prior distribution of vector β and we know the model's structure. In practice we usually know point estimates of β and additional information that helps us to evaluate how these estimates are reliable, for example expert's opinion of their reliability, discrepancy of experts' opinions, standard deviation of β, confidence interval of β. We could make different opinions on reliability of estimate for particular element β_i of vector β.

(2) resulting from our decision to create and then modify a model. In this case prior information is stored in data set we have access to. In the first step we represent prior knowledge in the form of prior model and in the next step we use this representation to create posterior estimations of the model parameters.

Having the point estimates of β and additional information that helps us to evaluate how these estimates are reliable, we also need to make some assumptions to choose its prior distribution. One of natural choices could be to make assumption that β have normal distribution $N(\mu,\sigma)$ with μ equal to point estimates of β and σ reciprocal to β reliability.

Someone could ask why not use data directly but confine oneself to the transformed information in form of a model? One reason was already mention – tackled problems caused by flat-maximum effect. Model representation could be also more convenient for combining different sources of information, especially when they are not compatible. In this case we use the model representation as a "bridge" spanning different sources of information increasing their compatibility. You may see case study at the end of the article for practical example.

Another reason is that data itself represents only the raw information whereas the model could contain also some other form of knowledge. The model could be created by incorporating the additional knowledge of an expert or the opinion of business party acquired during the process of model construction. This process of a priori model adjusting could take quite a long time and resources. Therefore we should value and use an existing scoring models if we have them available and they are performing well. We should be able to take benefits from them by using these models representation as the prior knowledge when constructing new models.

3. Baysian inference for linear regression model

In this section we assume that the available prior information is represented in a form of regression model with the vector of regression parameters $\beta = [\beta_1, \ldots, \beta_k]$, where k is the number of explanatory variables including the intercept. For better understanding how the method works we start with viewing classical linear regression model from the Bayesian perspective and then we describe how to add (informative) prior information to the classical model.

Classical Linear regression model from the Bayesian perspective

*B*asic model formulation

Linear regression model is formulated as follows:

$$y|\beta, \sigma, X \sim N(X\beta, \sigma^2 I) \qquad (1)$$

where n is the number of observations,
I is n x n identity matrix, X is matrix of explanatory variables,
β is the vector of parameters and σ^2 is a variance component.

Notice: that errors are independent and have equal variances. We will call this case the **ordinary linear regression**.

Standard noninformative prior distribution

Assuming standard noninformative prior distribution gives Bayesian parameters and error estimations that coincide with those produced by classical approach.
For normal regression model standard noninformative prior distribution is the uniform distribution of the parameter $\theta = (\beta, \log \sigma)$. It is the scale invariant distribution (for detail see Box, Tiao, 1973).

Bayesian model under the given parameter σ *and the regressor matrix X)* is represented as:

- Prior $\beta \sim U(\beta)$ *(improper uniform prior)* (2a)
- Data $Y|\beta \sim N(X\beta, \sigma^2 I)$ (2b)
- Posterior $\beta| \ Y \sim N(b, \sigma^2 V_\beta)$, (2c)

where: $b = (X^T X)^{-1} X^T Y$, (3a)
$V_\beta = (X^T X)^{-1}$ (3b)

Comparing this formulas with classical estimate (Greene 2000) we see that the estimate of β coefficient and covariance component are equal to those given by ordinary least square method. Classical (non-Bayesian) approach to modelling of linear regression could be thought as the special case of Bayesian posterior inference where prior distribution is assumed to be noninformative. It is easy to proceed from this perspective to fully Bayesian approach for linear regression.

Linear regression model with the prior information included.

We start with the remark that even in the classical regression approach (as described above) some prior information is already included. When we make assumptions about model functional form, variables included in the model etc., we could view these decisions as based on prior information we have. For example, excluding potential predictive variable X_i from the model could be viewed as imposing on parameter β_i the distribution concentrated in zero, that is the one with the expected value and variance equal zero.

To incorporate prior knowledge we need to move towards more general case of linear regression model that is formulated as follows:

$$Y | \beta, \Sigma_y, X \sim N(X\beta, \Sigma_y) \qquad (4)$$

where Σ_y is the covariance matrix of the error terms.

We assume here that errors are independent but have unequal variances. This case is called the **weighted linear regression**.

In this paragraph we assume that we possess prior information about the vector of regression parameter β. We also assume that prior distribution of β_i is $N(\mu_i,\sigma_i^2)$.

It can be easily proved that in this case we deal with conjugate family of distributions. Below we present two possible ways of posterior inference.

Simulation method:

Let us assume the scale invariant prior distribution for parameter $\theta = (\beta, \log \sigma)$, as previously.

To sample from the joint posterior distribution we need to:

1. Draw σ^2 from marginal posterior distribution
2. Draw β from posterior distribution under the drawn σ^2 in the previous step.

To carry out posterior predictive simulation we need to draw a random sample y^* from its posterior predictive distribution by the following procedure:

1. draw (β, σ^2) from their joint posterior distribution
2. draw $Y^* \sim N(X\beta, \sigma^2 I)$ (conditionally under the drawn parametersβ and σ^2 in the first step).

Example: For noninformative prior we have marginal posterior distribution for σ as follows

$p(\sigma^2|Y) = p(\beta,\sigma^2|Y)\,[p(\beta|\sigma^2,Y)]^{-1}$, (5)

Hence $\sigma^2|Y \sim Inv\text{-}\chi^2\,(n\text{-}k, s^2)$.[Gelman, et al., 1997] (6)

Then the posterior of the parameter β under the known σ^2equals (2c).

Before the simulation we need to calculate:b, V_β (see the formula 3a and 3b),

$s^2 = (Y - X\beta)^T (Y - X\beta)/(n-k)$, where k is the β dimension.

Such a simulation can be generalized to more complex models. However for practical usage of linear regression it is much more convenient to use augmentation method described below.

Augmentation method:

In this approach we benefit from the attributes of exponential family of distributions mentioned on page 3. We could interpret prior information about β_i as additional point of observation. Technically, we should append: (1) vector of $\boldsymbol{y}$ with additional $n + 1$ observation $y_{n+1} = \mu_i$,where i is a certain index (2) matrix X with one row filled with zeros except for the i-th element equal 1 (3) matrix of covariance Σ_y with one row filled with zeros except for the diagonal element equal σ_i^2. This step could be repeated for all other β_i parameters, so we could create vector $\mu = [\mu_1, \ldots, \mu_k]$, and covariance matrix Σ_β - with σ_i^2 on its diagonal to represent prior information about vector β.

The posterior distribution of β can be calculated by applying the method of weighted linear regression to the following augmented data:

$$y^* = \begin{pmatrix} y \\ \mu \end{pmatrix}, \quad X^* = \begin{pmatrix} X \\ I_k \end{pmatrix}, \quad \Sigma^* = \begin{pmatrix} \Sigma_y & 0 \\ 0 & \Sigma_\beta \end{pmatrix},$$

where I_k is the k dimensional identity matrix.

The β vector given by data (y^*, X^*, Σ^*) could be obtained by using wls.

4. Bayesian inference for nonparametric scoring model

There may be a case when the available a priori model is provided in form of decision rules, classification tree or segmented feature space with assigned classes. For the purpose of this chapter, we will consider all those models as the same type and use classification trees as representative example.

Formally, the problem of classification can be defined as follows. Observations are divided into separable segments described by possible variables' values. For example we can describe segment *S* as follows:

$$X_1 \in V_1^S \;\&\; X_2 \in V_2^S \;\&\; X_N \in V_N^S$$

Where V_i^S is a set of possible values for variable X_i in segment *S*. Each segment has assigned a probability distribution for particular decision classes. We can obtain this distribution from training data. Below we present an illustration of exemplary segment description:

$$age \subset (25;31) \;\&\; job \subset \{dentist;\ layer;\ football\ player\} \;\&\; average\_transaction \subset (1200;\ \infty)$$

During scoring we associate each data record with the correct segment using the segment description. Then we associate a new tested record with the probability distribution of decision classes that are assigned to selected segments.

The approach of upgrading classification tree models depends on the form of prior information we are constrained (or we are decided) to include. We could possibly have: data set used to create classification tree model; tree structure; distribution of the predictive variable within a segment; quantity of the observations within a segment; additional information from an expert about classification tree parameters etc. In practice we have some combination of those information. For example, we could have information about classification tree structure and all its parameters including quantity of observations in segments but without any additional information from an expert about classification tree parameters.

In our attempts to construct a new a posteriori model we can follow two main ways:

- We treat tree structure as given and we aim to find posterior target distribution by including new data within this structure.
- We try to find optimal classification tree for new data taking into account prior information.

Inferring about default probability for given classification tree structure

We assume that observations in all segments separately were sampled from a large population as the sequence of independent, exchangeable trials. Exchangeability means here that order of the sequence doesn't mater. For two level target we have binominal sampling model. Let *y* be the total number of ones (*1s)* (meaning "a true", "a success") in trial for one of the segments. We have

$$p(y|\theta) = Bin(y|n, \theta) = \binom{n}{y} \theta^{y} (1-\theta)^{n-y} \qquad (1)$$

Lets assume that θ have Beta prior distribution with parameters "*a*" and "*b*", θ ~ *Beta(a,b)*. From the definition of Beta distribution we have $P(\theta) \propto \theta^{a-1} (1-\theta)^{b-1}$ Posterior distribution of θ is proportional to the product of prior and likelihood (see page 3). That is:

$$P(\theta|y) \propto P(y|\theta)\, P(\theta) \propto \theta^{y}(1-\theta)^{n-y}\, \theta^{a-1} (1-\theta)^{b-1} \propto \theta^{y+a-1} (1-\theta)^{n+b-y-1} \qquad (2)$$

From the formula shown above you can notice that for Beta prior distribution and binominal distribution of *y* posterior distribution of θ is also *Beta(a + y, n + b - y)* that is the conjugate family for likelihood. We could use the benefits of conjugate family presented on page 3, namely we could use the possibility to represent prior information as additional data points.

For each segment *i* we find posterior distribution as $Beta(\alpha_i, \beta_i)$, where $\alpha_i = (a_i + y_i)$, $\beta_i = (n_i + b_i - y_i)$. From the properties of Beta distribution we have: $E(\theta) = \alpha_i / (\alpha_i + \beta_i)$. We can compute posterior probability of the target variable by adding to the *i*-th segment of the prior classification tree n_i observations with y_i observations that equals 1 and $(n_i - y_i)$ observations that equals 0.

Relaxing classification tree structure

We will present a way of how we can construct a new a posteriori model of a classification tree that may have different structure from the original a priori tree model. Such a situation may occur, for instance, in the case when a bank possesses a scoring card in the form of decision tree built for product A and wants to include this information into a new scoring card for product B in which some new variables are included. Therefore the new scoring model is expected to have a different structure than previous scoring card.

This method includes three steps: (1) transforming the feature space defined by the given tree model into a data table that represents prior information, (2) augmenting that table with a new data and (3) constructing a new tree based on joined data tables.

In the first step we extract the information from a priori tree model into data table format. Leafs of the classification tree define segments in feature space with assigned target value distribution, also called default probability within a segment. For each segment we generate a random data sample using conditional distribution defined by the leaf.

When generating the random samples covering particular segments, we should include information about variables' distributions, if it is available. An expert's knowledge can be taken into account to support this. In case of categorical variables (with limited number of discrete values e.g. credit purpose categories) an expert can assign the frequencies of variable values occurrence for the particular segment. If we don't have this information, an uniform distribution of variables' value is used. Dealing with continuous variables the expert should define a range of values of distribution (e.g. minimal and maximal value of variable such as age or income). Samples are usually generated from uniform distribution if we do not have any prior information about variable distribution.

Data samples generated by this process constitute so-called artificial training data set that covers the feature space described by the a priori model. All generated observations are assigned with a target value according to default target probability of the segment. The table is extended by column with α coefficient that will allow to determine a level of influence of a priori model in the a posteriori model.

Further process is rather simple. In the second step the data set containing new incoming observations is added to the table with generated artificial training data. Because these new data should be fully considered in the final model, they are assigned with an appropriate α value. As the third and last step, we construct a posteriori classification tree model using data from the joined tables. The classification model should be able to consider the weight parameter α of each data record in order to influence the significance of a priori model. The whole procedure allows us to obtain the final a posteriori model that takes into account information from a priori model when constructing new scoring models.

5. Baysian inference for logistic regression model

In this paragraph we consider parametric model aimed to predict a binary target variable. Although, we focus on logistic regression case, all consideration presented here could be easily extended for other binary regression models as well.

We assume that model is given by the equation:

$$y_i \mid p_i \sim Bin(m, p_i),\ m=1 \qquad (1)$$

where $p_i = Logit^{-1}(X\beta)$, $Bin(p_i)$ stands for a Bernoulli distribution with parameter p_i, $logit^{-1}(u) = [u/(1-u)]$.

Let us assume that prior information about logistic regression parameter β_i is represented in the form of normal distribution:

$$\beta_i \sim N(\beta_{0ii}, \sigma_i^2) \qquad (2)$$

For the logistic regression model the likelihood is given by:

$$p(y|\beta) = \prod_{i=1}^{n} \left(Logit^{-1}(X\beta)\right)^{y_i} \left(1 - Logit^{-1}(X\beta)\right)^{1-y_i} \qquad (3)$$

The posterior distribution is proportional to the product of prior distribution and likelihood:

$$p(\beta|y) \propto p(y|\beta) \cdot p(\beta)$$

Following (2) and (3) we have:

$$p(\beta|y) \propto \prod_{i=1}^{n} \left(Logit^{-1}(X\beta)\right)^{y_i} \left(1 - Logit^{-1}(X\beta)\right)^{1-y_i} \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi\ \sigma_j^2}} \exp\left(-\frac{(\beta_j - \beta_{0j})^2}{2\sigma_j^2}\right)$$

The distribution $p(\beta|y)$ defined above does not represent any form of posterior distribution that with the prior distribution given by (1) and likelihood given by (2) would belong to conjugate family of distribution. There isn't any natural candidate for prior distribution that allows us to use benefits of conjugate family. Especially, we can not represent prior information as additional observations that augmented those used to estimate model parameter β, in the way it was presented for linear regressions and classification trees. In such a situation Monte Carlo simulation is a right tool to use. In this approach we generate a Markov chain with stationary distribution equals to posterior distribution of vector β we are looking for [Tiermey, 1994].

Markov-chain-based dynamic Monte Carlo method was developed in 1950s by statistical physicists N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller. In the paper of Metropolis (1953) a new method called **Metropolis algorithm** was introduced. Later, in 1983 it was used to develop a **simulated annealing** technique (Kirkpatrick, Gelatt, Vecchi) applicable in solving optimization problems. Currently, Markov-chain-based methods are used in wide range of areas such as biology, economics etc[Bremaud, 1991].

The idea of Markov chain simulation is to generate a chain of random points in the space of θ where distribution converges to posterior distribution $p(\theta|y)$ that we are searching for[Tiernet, 1994]. The points of Markov chain have Markov property. This property means that only the present state gives any information of the future behaviour of the process. The key element of Markov chain is transition distribution $T(\theta_{t+1} \mid \theta_t)$ for which we draw θ_{t+1} dependently on θ_t .

The construction of transition distribution should guarantee the chain to be ergodic and then the chain has its unique stationary distribution that is identical with posterior distribution $p(\theta|y)$ that we are searching for. To speed up the convergence, transition distribution $T(\theta_{t+1} \mid \theta_t)$ could be constructed as function dependent of the simulation's step. The simulation of θ_t, which leads to a stationary and unique distribution, not necessarily should be based on

Markov property but assuming those properties considerably eases the proof of convergence for broad class of transition distributions.

It is worth mentioning that in traditional Markov chain analysis we know the transition rule and we are interested to find what stationary distribution is. In simulation of Markov chain we know the equilibrium distribution and what we are looking for is the efficient transition rule that guides the chain to the equilibrium. The Metropolis algorithm defines transition rule for Markov chain. It enforces symmetric jumping distribution (the key element of transition function $T(\theta_{t+1} | \theta_t)$). Metropolis algorithm was further extended by W.K. Hastings. The main modification was to allow for asymmetric jumping distribution. We used Hastings-Metropolis algorithm in our simulation as is described later.

6. Case Study – Scoring model for credit payment

In this chapter we will present several experiments made with the usage of real data to illustrate the process of updating provided priori model by new data. Data preprocessing and modelling was done in software called Górnik System. It is a flexible tool for data analysis and mining provided by StatConsulting Ltd. For the experiments we used modules for logistic regression algorithms, Monte Carlo simulations and general statistics libraries. More information about the software can be found on http://www.StatConsulting.biz.

Business Task

The business task was to predict the risk of the company's customer not paying the granted credit for specific product. Based on experts' opinion a new procedure was introduced with the aim to enrich the sources of information that the acceptance decision would be based on. Customers were obliged to fulfill extended application form with new questions and submit additional documents which referred to their employment. The consequence was an increase of the number of predictive variables which could be included in the model.

Modelling team had two available sources of information: (1) Existing scoring model based on sound data of over one hundred thousand records, but without new introduced variables (2) New incoming records of clients treated with new procedure.

Traditional approach of scoring models construction leaves only one option – to use a generic scoring model until there is enough amount of data with extended list of variables and then to create a new model based solely on new records. It was a good opportunity to use the Bayesian approach in order to construct a new model that would take into account the changes of the scoring procedure. The description of data preparation and analytical solution is based on simplified business problem due to the legibility of presentation. The problem of handling rejected applications was treated as a separate business task and is not discussed in this article. The logistic regression model was supplied by the output of decision tree(s) modelling (leaves of the tree were used as predictive variables). The description of this process was also omitted in this article.

Data description

The sources of information used to prepare predictive variables were based on:

I. Demographic information about customer, for example:

- customer's age when fulfilling application form,
- customer's gender,

II. Demographic information about region of customer residence, for example:

- level of "bad" payers in the region of customer's residence,

- concentration of entrepreneurs in the region of customer's residence,
- population in the region of customer's residence,

III. Information gathered during a verification process,

IV. Information about customer's choice concerning the offer and the product.

The generic scoring model used 37 attributes labelled later in the text as $X_1, \ldots X_{37}$. Introducing new procedure allowed to extend predictive variables set with additional two source variables which after pre-processing gave 8 binary variables labelled as $X_{38}, \ldots, X_{45}$.

Analytical solution

Data set preparation

We constructed logistic regression models to predict probability that customer will fulfill his financial obligation.

For four subsequent periods of time we carried out the following procedures:

- generic model modification based on new data. We denoted M_{Ai} as the posterior model based on new data available in period i (i = 1, . . . , 4),
- model construction based only on new data. We denoted M_{Ni} as the model based only on new data available in period i (i = 1, . . . , 4),
- validation of generic model M_G against M_{Ai} and M_{Ni} based on test data. To make validation more reliable for the test data we included customers that were granted a credit after all periods. The group of 16 560 customers was used for the test.

For modelling we used weighted sample of 70% of "good" and 30% of "bad" customers. The number of observations in each period is presented in table1:

	Number of observations
Period 1	422
Period 2	835
Period 3	1724
Period 4	6806
Test data	16560

Table 1. Quantity of observations used for modelling and testing

The results of the variants' comparison we present further in this chapter.

Setting priors

Let β be the vector of parameters of interest (we search for posterior distribution of β). We set prior distribution β_{0i} of chosen element β_i of vector β as $N(\beta_{iA}, \sigma_{iA})$, where β_{iA} and σ_{iA} are, respectively, expectation and standard deviation of β_{iA} in the model M_G. This means that we use M_G as the source of prior information about β.

Choosing starting point for simulation

This is a crucial step for Markov chains simulation. Choosing wrong starting point for generating chains could cause the sampling from the area where probability is almost zero. It is easy to put oneself in such uncomfortable situation due to the fact that for the most distributions entire mass of probability is concentrated near a few modes (one mode in our case). For example for normal distribution probability density is drastically diminishing when we move several standard distributions from the mode. To overcome this difficulty we chose starting point for beta search as the $\beta_0 = \alpha\beta_A + (1 - \alpha)\beta_B$ proportionally to the ratio between standard deviation of these parameters. The motivation is that posterior beta is created as modification of prior distribution towards sample distribution.

Setting the Hastings-Metropolis algorithm

The main parameters that we need to control during the simulation are: (1) the number of chains, (2) the number of observations in a chain – we will make inference based on simulated sample, we have to trade-off between precision and computing time (i.e. the time we wait for result), (3) the cut-off for the chain – the beginnings of the chains are not stationary. We have to cut the observations and leave only those that are stationary according to predefined measure.

The major problems we encounter during simulation refers to: (1) finding whether we are close enough to the search posterior distribution, (2) omitting correlation between subsequent points in Markov chain. We used square root of R statistics to control convergence of the chain. The idea is that chain reaches stable state when the variance between parallel chains is approximately the same as the variance within the same chain when the simulations are close to stationary distribution. R statistics is given by: , where *B* and *W* is between and within variance of the sequence of Markov chain, respectively (see Liu 2001 for details). To tackle the problem with correlation between subsequent points in Markov chain we tried to increase the number of generated chains (up to several thousands) and we used only one observation from the chain. Alternative approach was to pick from the sequence only some observations and discard the other points in Markov sequence.

Results of simulation

All presented results are based on test data set. Posterior β distribution charts against prior distribution for chosen variables X_1 to X_4 are presented in appendix 1.

In appendix 2 the accuracy statistic was plotted against the size of training sample. Accuracy shows the ratio between correctly and incorrectly classified observations calculated for given threshold of posterior probability. In appendix 2 we show results for threshold equal 0.5. For this threshold the accuracy equal 0.8 means that 80% of the customers with higher probability of being "bad" than "good" are correctly classified. The results are shown for three models: generic model, model with new data only and model with prior information included. Figure 1 – 3 reflects different levels of parameter alpha. We used the alpha parameter to control the influence of prior information on posterior distribution. The prior variance of model parameters is multiplied by the parameter alpha.

From Figure 1 – 3 we may notice that accuracy statistic for model with prior information included is better for period 1 – 3 with sample size equal 422, 835, 1724

observations respectively. The result is consistent for all levels of beta parameter. For period 4 with 6802 new observations model with prior information gives the best result only for alpha equal 10. It means that the influence of prior information should be tuned to fit both reliability of prior and sample information. For smaller data set we notice that the impact of the prior information is more crucial for predictive accuracy of the model. **The final results show that including prior information produce better predictive power measured by accuracy statistic.** Analogical results based on lift statistic with alpha equal 1 are shown in appendix 3. For convenience of the readers we presented in appendixes only some representative results. More information can be found on www.statconsulting.biz\edinburgh2005 .

7. Conclusions

This work is a result of StatConsulting's research and implementations of our analytical solutions concerning the problem of generic scoring models adjustments to new specific situations. We received such requests from several different financial institutions towards our company. That indicates that there exists a significant need for such solutions, where knowledge hidden in the existing scoring models could not be ignored in the process of constructing a new model.

In the paper we presented several ways how we can update scoring models by new information. This can be converted into the problem of constructing empirical scoring models that would consider two sources of information: a priori knowledge in form of generic model and inflowing data about behaviour of new customers. We have provided theoretical introduction to Bayesian approach used in this kind of problems and we analysed several cases of using different types of models: classification trees, linear and logistic regression.

In the case of linear regression and classification tree with fixed tree structure we can straightforward incorporate prior knowledge based on properties of conjugate family of distribution. Logistic regression case requires more advanced approach based on Monte Carlo simulation. Both approaches differs in computational complexity and easiness in reaching posterior estimation of searched parameters.Presented results show that:

(1) It is computationally possible to update scoring model in form of logistic regression by finding posterior distribution of its parameters based on Markov Chains simulation for real data problems encountered in credit scoring area,

(2) incorporating prior knowledge could improve predictive power of the model.

We continue the research in this area and study the usage of described methods in business practise. If you are interested in the subject, feel free to contact us.

8. References

1. Berger, J. Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, 1985
2. Box, G.E.P, and Tiao, G.C. (1973) Bayesian Inference in Statistical Analysis, Addison-Wesley, Reading, Massachusetts.
3. Bremaud, P., Markov Chains, Gibbs Fields, Monte Carlo Simulation and Queues, Springer, 1991, New York.
4. De Groot, M. Optimal Statistical Decisions. McGraw-Hill, New York, 1970
5. Diaconis, P., Ylvisaker, D. (1985), Conjugate priors for exponential families. Ann. Statist., 7, 269-281.
6. Gelman, A, B., Carlin, J.S., Stern, H.,S., Rubin, D.,B. Bayesian Data Analysis, Chapman & Hall, 1997, New York.
7. Green WH, Econometric Analysis, Prentice-Hall International, 2000
8. Kirkpatrick, S., Gelatt, C., Vecchi, M., Optimization by simulating annealing, Science 220: 671-680
9. Liu, J.S. (2001). Monte Carlo Strategies in Scientific Computing, Springer-Verlag New York, Inc
10. Lucas Alan, Powell Joanna, "*Small sample scoring*", Proceedings of Credit Scoring and Credit Control VII, Credit Research Centre, University of Edinburgh, 1997.
11. Metropolis, N. and Rosenbluth, A. and Rosenbluth, R. and Teller, A. and Teller, E. (1953) Equation of state calculations by fast computing machines J. Chem. Phys. 21:1087-1092
12. Robert, C.,P. The Bayesian Choice, Springer, 1994, New York.
13. Savage LJ, The Foundations of Statistics, John Wiley & Sons, New York , (1954)
14. Stamokostas Konstantinos, Vasiliou Dimitrios, Adraktas Georgios "*Risk-Based Pricing (RBP) Using Bayesian Statistics: How to Market RBP in the Context of New Credit Card Customers*", Proceedings of Credit Scoring and Credit Control VII, Credit Research Centre, University of Edinburgh, 2003.
15. Tierney, L. (1994), Markov Chains for Exploring Posterior Distributions, Ann. Statist., 22, 1701-1762.

Appendix 1 Posterior β distribution against prior distribution

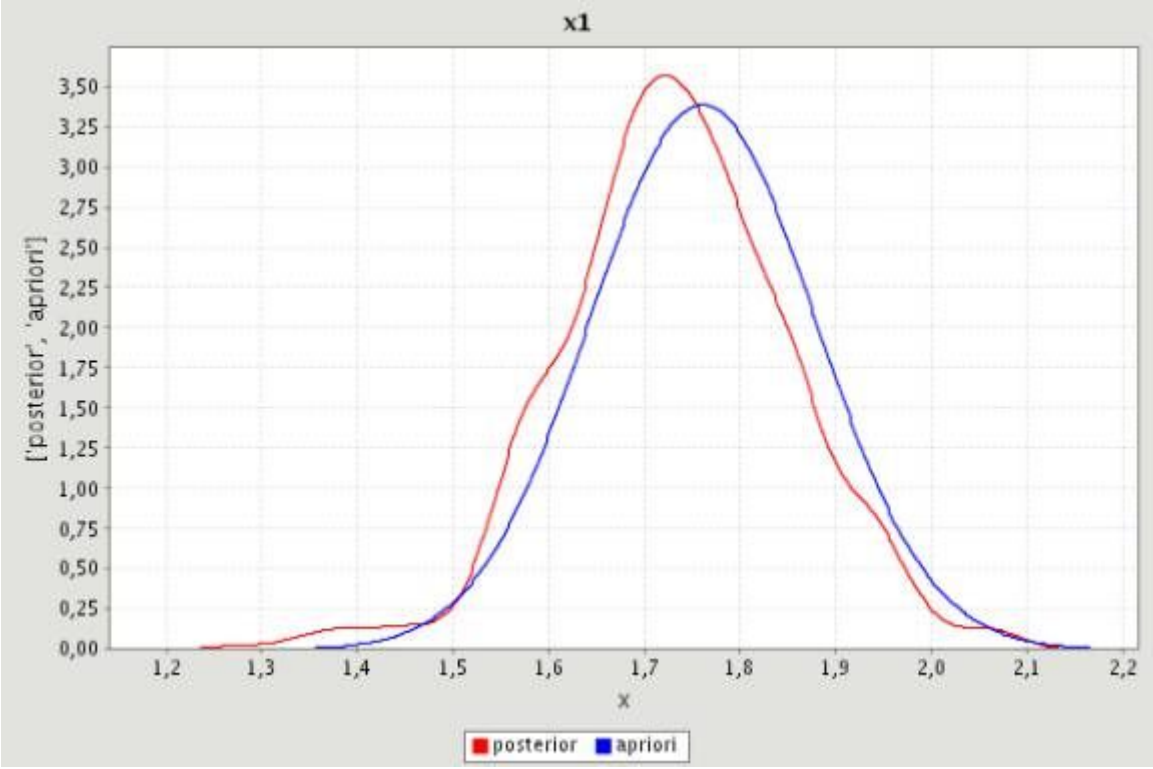

Figure 1. Posterior β distribution against prior distribution for variable X1

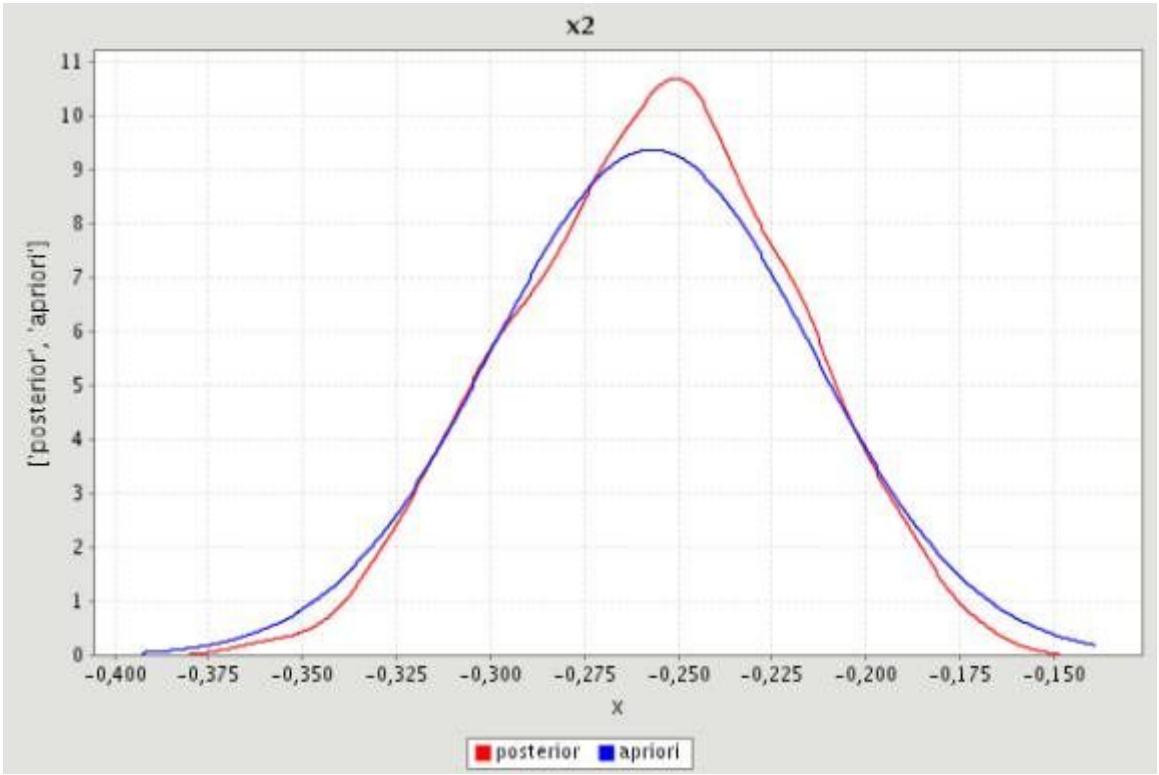

Figure 2. Posterior β distribution against prior distribution for variable X2

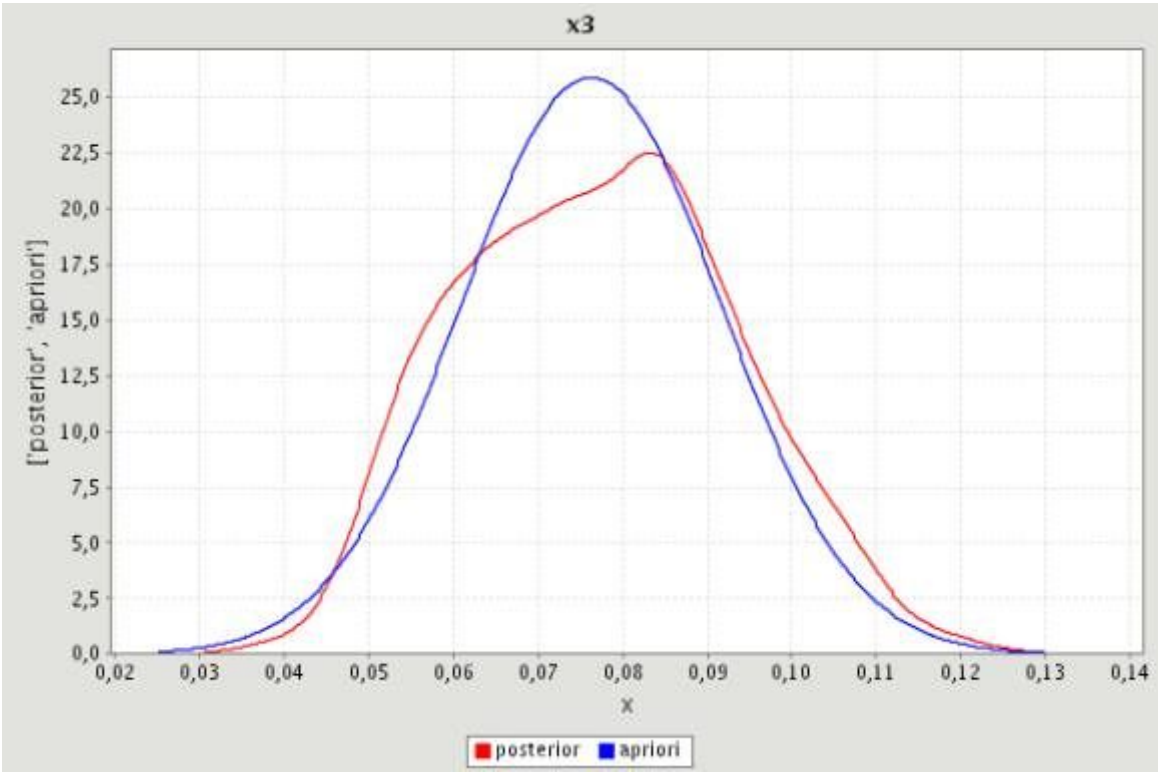

Figure 3. Posterior β distribution against prior distribution for variable X3

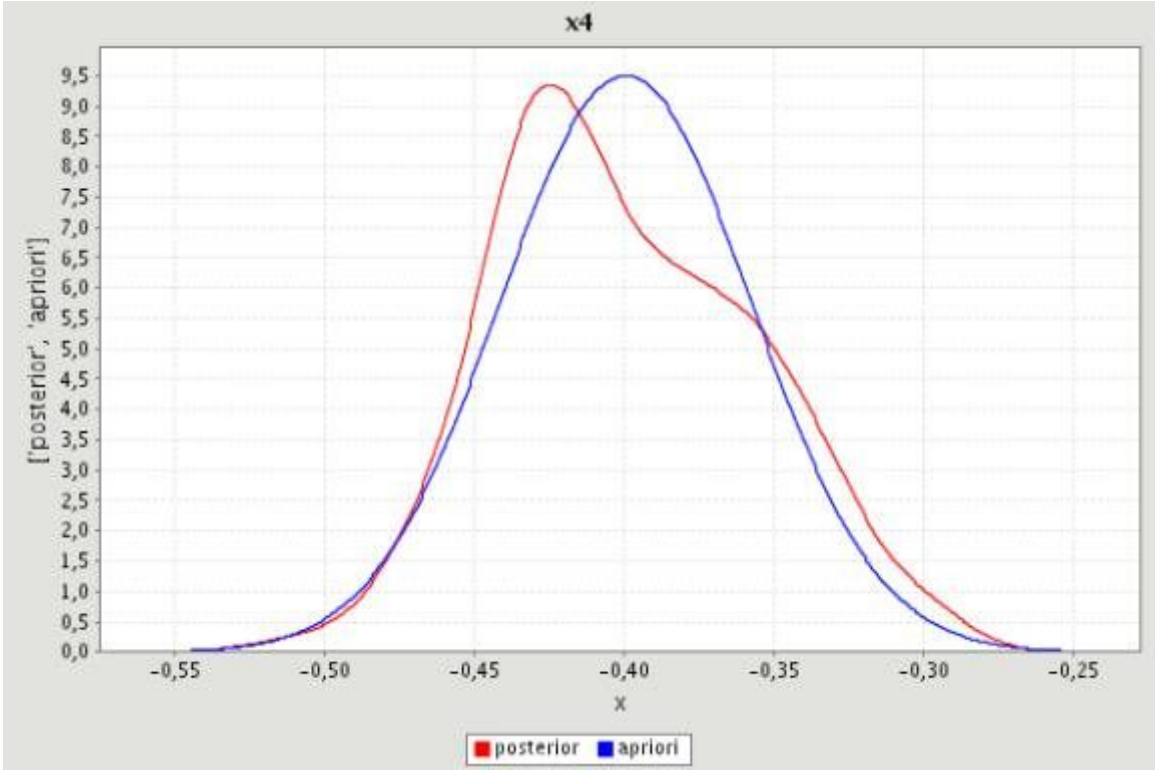

Figure 4. Posterior β distribution against prior distribution for variable X4

Appendix 2 Accuracy vs sample size

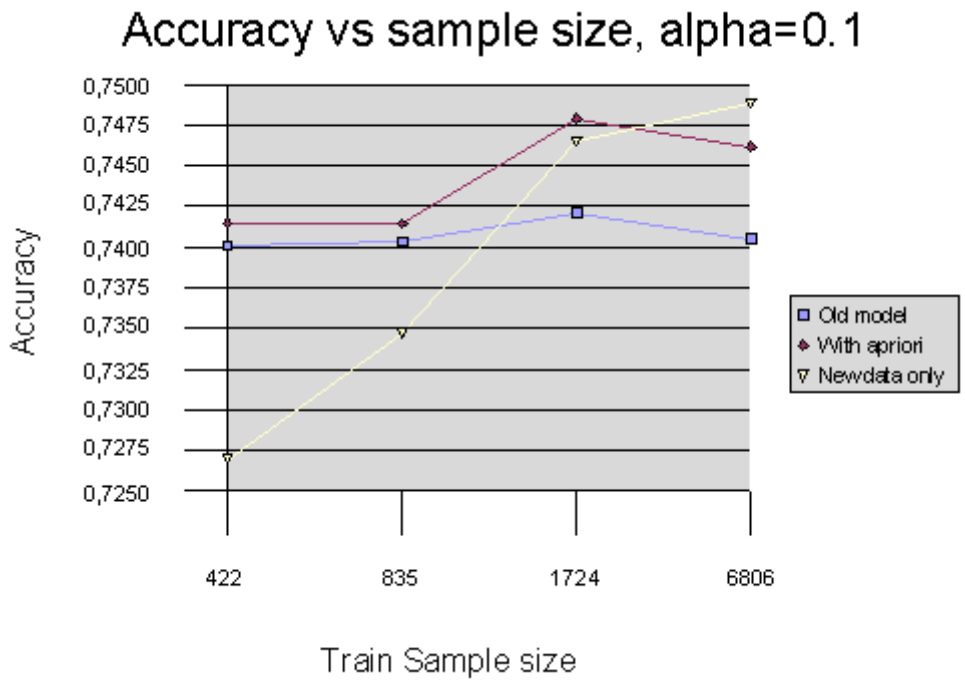

Figure 1. Accuracy vs sample size for parameter alpha equal 0.1

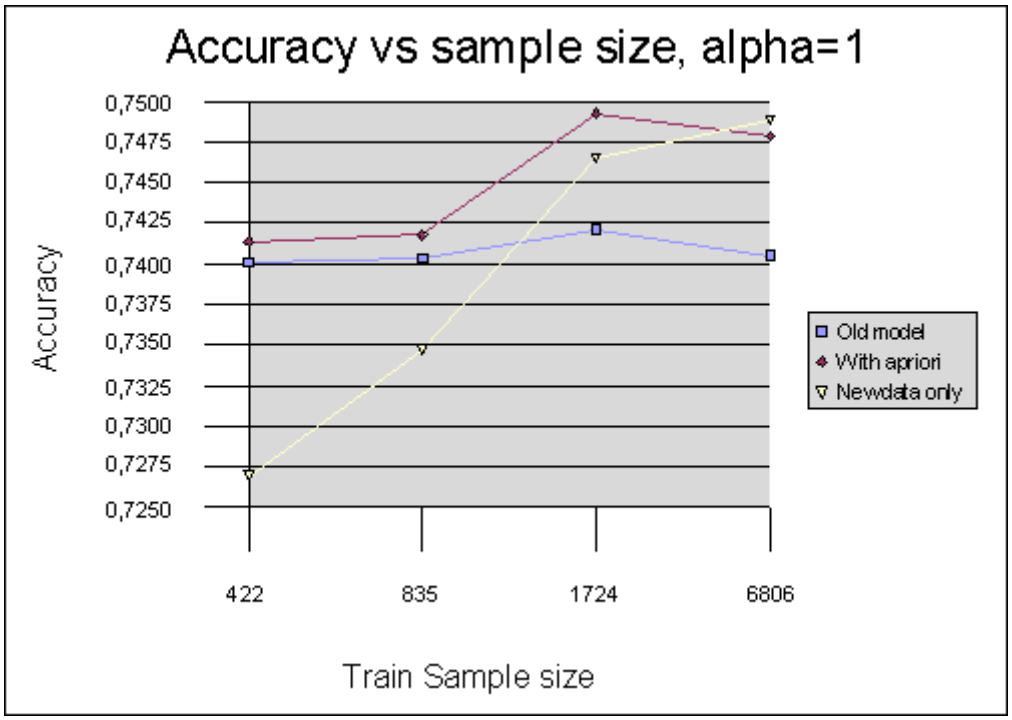

Figure 2. Accuracy vs sample size for parameter alpha equal 1

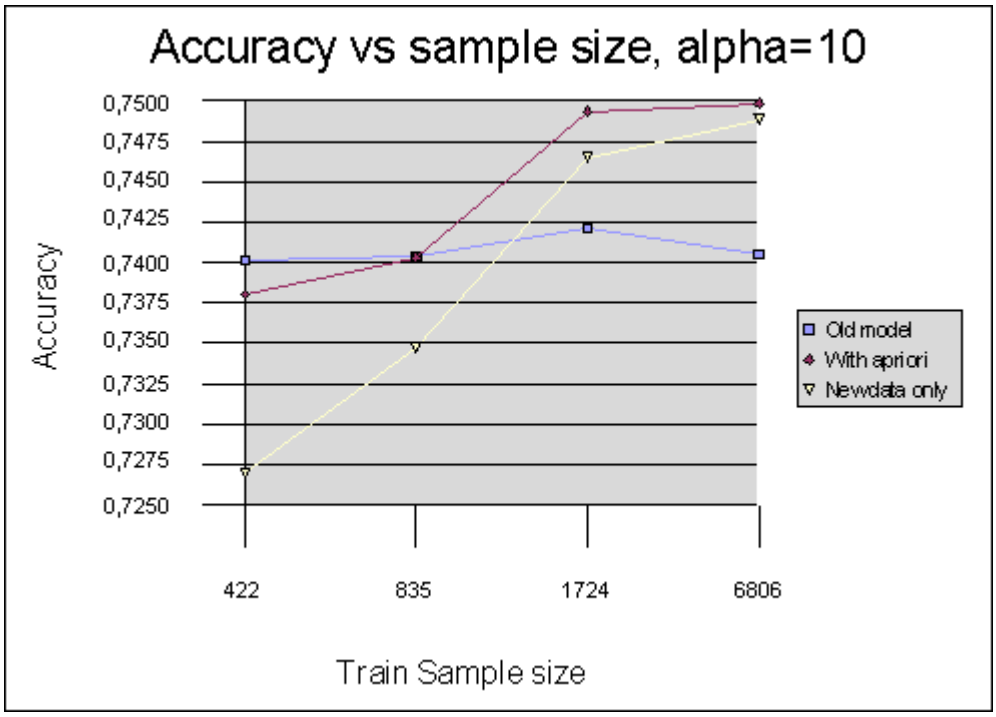

Figure 3. Accuracy vs sample size for parameter alpha equal 10

Appendix 3 Lift charts with alpha=1and different size of new data set

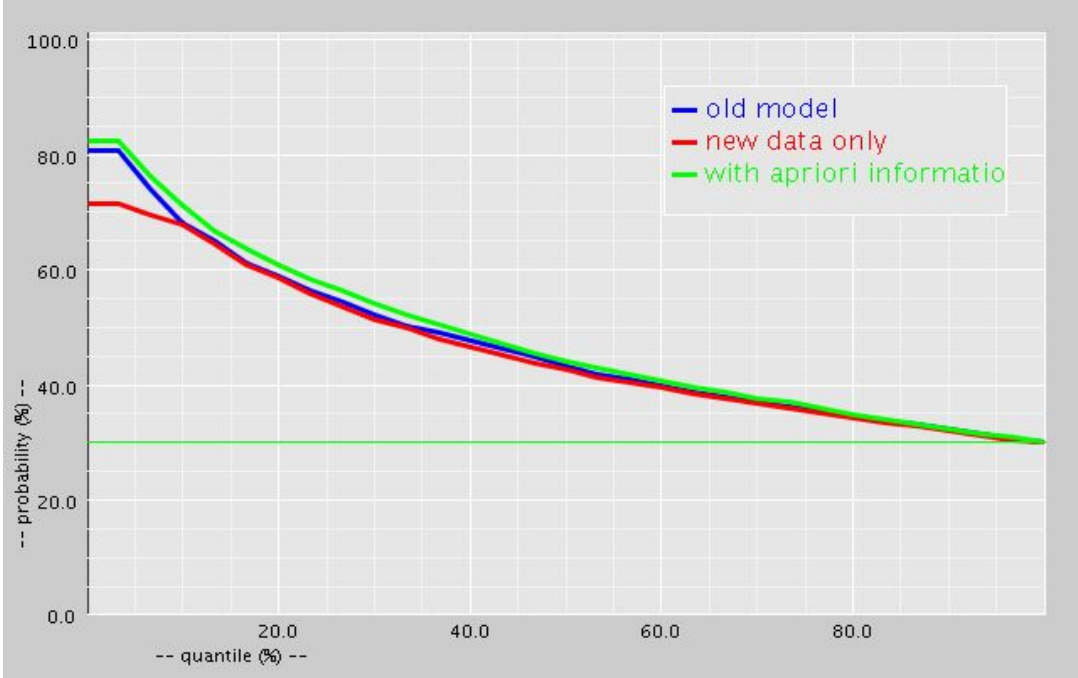

Figure 1. Lift charts for size of new data set equal 422 observations

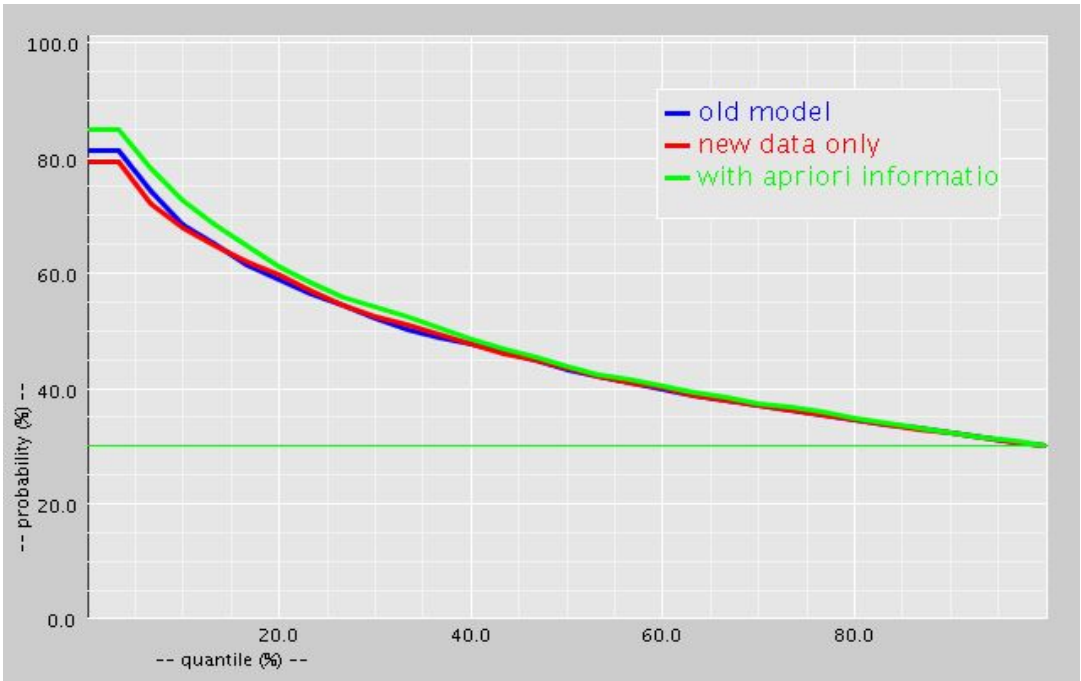

Figure 2. Lift charts for size of new data set equal 835 observations

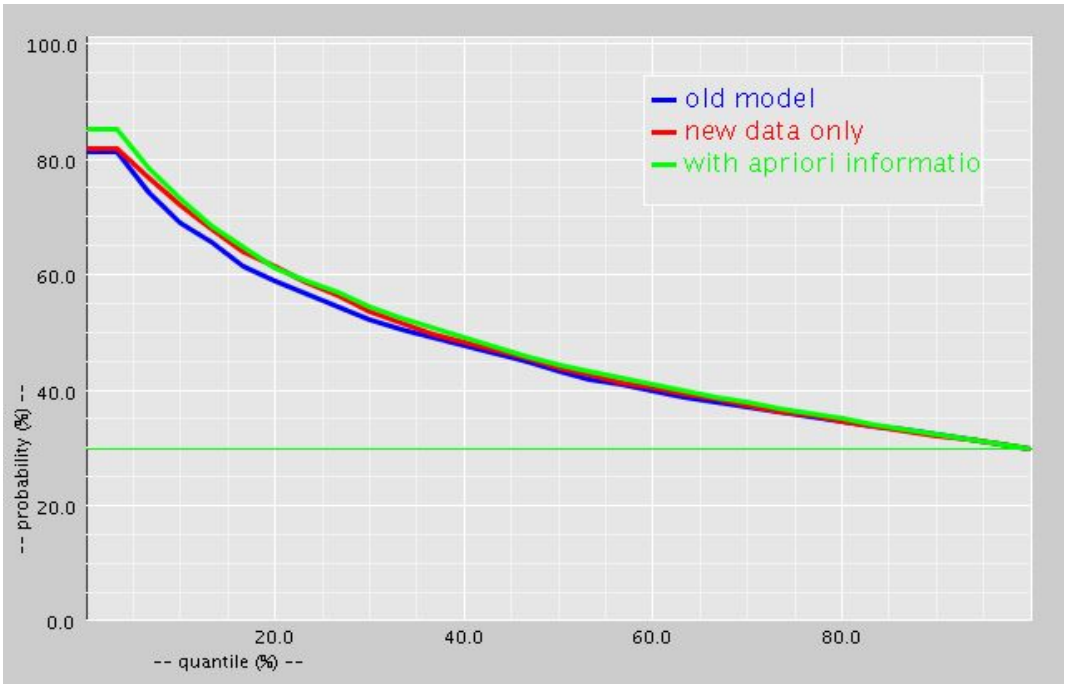

Figure 3. Lift charts for size of new data set equal 1724 observations

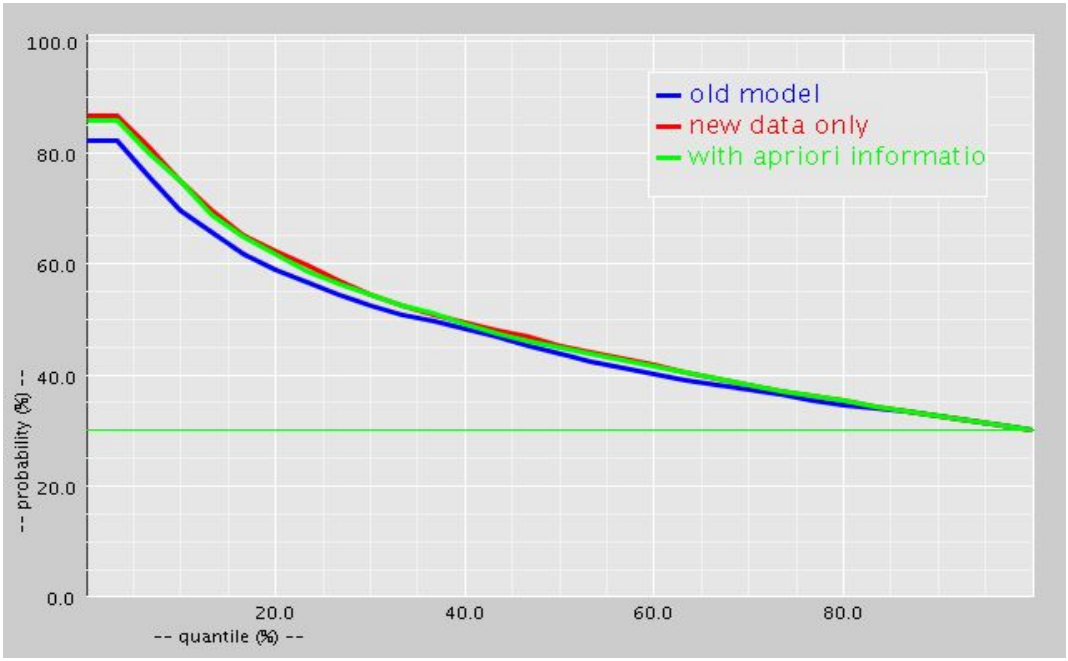

Figure 4. Lift charts for size of new data set equal 6806 observations