Kalman filtering as a performance monitoring technique for a propensity scorecard

Katarzyna Bijak*
University of Southampton, Southampton, UK, and Biuro Informacji Kredytowej S.A., Warsaw, Poland

Abstract

Propensity scorecards allow forecasting, which bank customers would like to be granted new credits in the near future, through assessing their willingness to apply for new loans. Kalman filtering can help to monitor scorecard performance. Data from successive months are used to update the baseline model. The updated scorecard is the output of the Kalman filter. There is no assumption concerning the scoring model specification and no specific estimation method is presupposed. Thus, the estimator covariance is derived from the bootstrap. The focus is on a relationship between the score and the natural logarithm of the odds for that score, which is used to determine a customer's propensity level. The propensity levels corresponding to the baseline and updated scores are compared. That comparison allows for monitoring whether the scorecard is still up-to-date in terms of assigning the odds. The presented technique is illustrated with an example of a propensity scorecard developed on the basis of credit bureau data.

Keywords

Propensity scorecard, scorecard monitoring, Kalman filtering, bootstrap.

Introduction

Propensity scoring

According to Thomas *et al* (2002, p 1), credit scoring is "the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit". Nowadays most banks use scoring to predict the credit risk of their actual or potential customers. Scoring

* I am grateful to an anonymous referee for suggestions and comments that helped improve this paper. I would also like to say thank you to Sophie N'Jai for proof-reading the draft.

models are also developed by credit bureaus to help banks assess credit risk on the basis of data coming from the banking sector as a whole. The most common form of such models is a scorecard. Mays (2004, p 63) defines the scorecard as "a formula for assigning points to applicant characteristics in order to derive a numeric value that reflects how likely a borrower is, relative to other individuals, to experience a given event or perform a given action". The characteristics (variables) can have several discrete attributes to which the scorecard assigns points (attribute scores). A customer's score is calculated as a sum of the attribute scores.

Most credit scoring models and techniques can be adapted for other bank activities such as collection (see Mays, 2004, p 7), fraud detection or marketing (see Thomas *et al*, 2002, p 4). In order to select customers for marketing campaigns (especially direct-mail ones), some banks use propensity scorecards that allow for the forecasting of which of their customers will soon be interested in new credits. Such models facilitate the prediction of customers' willingness to apply for new loans (credit propensity) in the same way that the credit scorecards make it possible to predict credit risk. While usually the higher the credit score, the lower the risk (and the better the customer), it is assumed here that the higher the propensity score, the lower the customer's willingness to apply for new loans. However, in practice, propensity scorecards are sometimes scaled so that the higher the score, the higher the willingness (and the more attractive the customer).

In credit scoring, customers are divided into goods (creditworthy) and bads (uncreditworthy). Similarly, in propensity scoring they can be divided into the willing and the unwilling to apply for new loans. In this research the willing customers are defined as those who applied for new loans in a four-month outcome period between the observation point and the outcome point. The observation point is a date on which data on a customer's behaviour are collected, and the outcome point refers to the date on which their status is determined. Because the credit bureau data, that pertain to the whole banking sector, are used, a customer's status is determined regardless of which bank they applied to for a loan.

A ratio of goods to bads is referred to as the odds in credit scoring. Similarly, here the odds are defined as a ratio of the unwilling to the willing among customers having a given score or a score coming from a given range. In particular, the odds can be calculated as a ratio of the unwilling to the willing in the whole sample. Irrespective of whether the odds are computed for a score, a score range or a sample, they can be treated as a measure of credit propensity.

Scorecard monitoring

Once a scorecard has been implemented, its monitoring (usually called validation) has to be performed regularly. According to Thomas *et al* (2002, p 17), "monitoring a scorecard is a set of activities involved in examining the current batch of applications and new accounts and assessing how close they are to some benchmark", which is usually determined on the basis of the development sample. A distinction is made between monitoring and tracking; the latter consisting of comparing expected and observed performance of cohorts of accounts over time. However, other researchers consider tracking reports as a specific type of monitoring reports.

A complete set of scorecard monitoring reports is suggested and described in detail in Mays (2004, chapter 13). Those reports can be divided into front-end and back-end ones. Front-end reports do not require information about defaults. There are reports on population stability and approval rate, characteristic analysis, override rate and override reasons. On the contrary, back-end reports are based on information about defaults. There are good/bad separation and early performance score reports. The good/bad separation reports allow for the evaluation of how well the scorecard separates goods from bads (using e.g. the Kolmogorov-Smirnov (KS) statistic). They also enable the observation of changes in a relationship between the score and the odds for that score. Those shifts - as well as changes in discriminatory power - can be identified by comparing current results with previous ones (e.g. one year ago) and with those based on the development sample. In order to produce a good/bad separation report, one has to collect data covering an outcome period of the same length as assumed in the scorecard development process (which is usually at least twelve months in credit scoring). Since this takes some time, the early performance score reports can be useful. Those reports consist of bad rates in established score ranges. There are bad rates in different cohorts of accounts after, for example, three and six months of booking. For even earlier evaluation of the scorecard effectiveness, the default can be replaced with 30+ or 60+ days-past-due in those analyses (see Mays, 2004).

Another set of monitoring reports is proposed and described in Anderson (2007, chapter 25). There are the following report types: portfolio analysis, performance monitoring, drift report, decision process monitoring and others (override analysis etc.). Portfolio analyses include delinquency distributions and transition matrices, while drift reports cover population stability checks and possible score shifts. Performance monitoring consists of examining

discriminatory power, accuracy (calibration) and stability of the scorecard. There are scorecard performance reports, vintage analyses and score misalignment reports. Scorecard performance reports are similar to the good/bad separation reports, while vintage analyses are similar to the early performance score reports. The score misalignment reports allow for the identification of problems at the characteristic level: points assigned to attributes of one or more characteristics might have stopped reflecting credit risk related to those attributes (see Anderson, 2007).

Standard scorecard monitoring reports are also described in Siddiqi (2006, chapter 9); the usual methods of scorecard monitoring are presented in Lucas (2004) as well as in Van Gestel and Baesens (2009, p 269-272), and some useful advice on the topic is provided in Schiffman (2001). Moreover, detailed information on measuring different aspects of scorecard quality (including a wide selection of discriminatory power measures) can be found in Thomas (2009, chapter 2).

The above-mentioned reports are designed for credit scoring models. However, most of them can be used to monitor scorecards applied in other areas. Obviously, the default has then to be replaced with the modelled phenomenon, and the customer's status has to be redefined accordingly. In particular, performance reports can be prepared for a propensity scorecard.

The main drawback of the commonly used approach to such reports lies in using - besides the development sample - only the current monitoring sample which is collected for one selected moment and thus may be atypical (e.g. because there was a period in which some credit products have been offered at unusually attractive conditions). Whittaker *et al* (2007) present a new scorecard performance monitoring technique that is free from the above-mentioned disadvantage. The technique is derived from the Kalman filtering. There is an assumption that the model parameters change constantly and the successive monitoring samples provide their measurements. Those measurements are used to update the baseline model and the updated scorecard is the output of the Kalman filter. The technique is demonstrated for a logistic regression model estimated using the maximum likelihood method, and illustrated with an example of a dynamic mortgage scorecard (see Whittaker *et al*, 2007).

In this paper the same technique is used but a more general approach is presented and applied. There is no assumption concerning the scoring model specification and no specific estimation

method is presupposed. Unlike in Whittaker *et al* (2007), tracking all attribute scores is not of interest here. The focus is on a relationship between the score and the natural logarithm of the odds for that score. That relationship is used to determine the propensity level of a customer having a given score. The log odds estimate, which represents the propensity level (provided that the baseline scorecard is still up-to-date), is compared with the estimate calculated using the relationship between the updated score and the log odds. That comparison allows for controlling whether the scorecard is in fact up-to-date in terms of assigning the odds. As an example a propensity scorecard is used, developed and systematically updated on the basis of credit bureau data.

The presented model is a sample one, developed only for the purposes of this research, and the analysed propensity scores differ from those offered by the Polish credit bureau, Biuro Informacji Kredytowej S.A. (BIK).

Methodology

Kalman filter

The Kalman filter is a common method for estimating the state of a noisy process (see Kalman, 1960). It enables the estimation, when the exact state cannot be observed and there are only some measurements (observations) which contain a noise. The method allows "filtering" the measurements in order to remove that noise (see Wells, 1996, chapter 4). It is assumed that the current state of a process depends stochastically on the previous state. This relationship is described by the state equation (also known as the transition equation). It is also assumed that the measurement depends stochastically on the state at the same moment. That relationship is described by the observation equation (also known as the measurement equation). Thus, the Kalman filter is used to estimate the state of a process governed by the state equation, when a link between the measurement and the state is expressed by the observation equation.

The above-mentioned equations create the state space model (see Harvey, 1990, chapter 3). According to Welch and Bishop (2006), the Kalman filter is "a set of mathematical equations that provides an efficient computational (recursive) means to estimate the state of a process". There are two groups of equations, which allow the Kalman filter estimates to be calculated:

time update ones and measurement update ones. The results of the time update equations constitute *a priori* estimates, which are then used, together with measurements, in the measurement update equations to obtain *a posteriori* estimates (see Welch and Bishop, 2006). The *a posteriori* estimate of the state of a process is an output of the Kalman filter.

In particular, parameters of a statistical model can be treated as the state of a process (since they are likely to change over time). It seems reasonable to assume that there is a stochastic dependence between them now and in the past. Obviously, it is not possible to calculate the exact values of such parameters. However, they can be estimated using an appropriate estimation method (e.g. maximum likelihood, ML) and their estimates can be thought of as the measurement. Thus, there must be a relationship between the measurement and the model parameters.

The measurement is assumed to contain a noise (e.g. as a result of sample selection). Therefore, it has to be "filtered" in order to determine the actual estimates of the model parameters. In this case, the *a priori* estimate of the current state is the previous output of the Kalman filter. Then, the *a priori* estimates are updated into the *a posteriori* ones using the measurement (e.g. MLEs). The *a posteriori* estimates of the model parameters are treated as the actual ones and constitute the Kalman filter output. Hence, the (actual) current estimates depend both on the previous ones and on the measurement.

As suggested in Whittaker *et al* (2007), in this research the scorecard parameters are the state of a process. Their estimation, which is based on a monitoring sample, provides the measurement. The actual estimates of the scorecard parameters are obtained using the Kalman filter. The output of the Kalman filter is referred to as the updated scorecard. The starting model, which is estimated on the basis of the training dataset, is called the baseline scorecard.

Baseline scorecard

The baseline scorecard is developed using a random sample S_0. The sample S_0 is randomly divided into training and test datasets that include, for example, 60% and 40% of customers, respectively. In both the datasets the same odds are ensured. All variables, which describe a customer's behaviour, are binned and then some of them are selected into the scorecard. The binned variables are used in the form of dummies. The model parameters are estimated on the

basis of the training dataset and the scorecard discriminatory power is confirmed on the basis of the test dataset. As a result there are the baseline model parameter estimates $\hat{\beta}_0^b$ which then are used as initial *a priori* estimates in the Kalman filter.

State equation

The state of a process is constituted by the scorecard parameters β_t, while the state equation describes the relationship between the current state and the previous one. It is assumed that this relationship takes the form of a multi-dimensional random walk:

$$\beta_t = \beta_{t-1} + q_t ,$$

where $\text{var}(q_t) = Q$ for all *t*. According to Whittaker *et al* (2007), it is assumed that the covariance matrix *Q* does not depend on time and that the individual model parameters vary independently. As a result of the latter assumption, off-diagonal entries of the matrix *Q* equal zero. Because all variables, which are used in the model, are in the form of dummies, diagonal entries of the matrix *Q* are equal. In consequence, the matrix *Q* is a diagonal one:

$$Q = \sigma I ,$$

where σ is referred to as a signal to noise ratio (see Whittaker *et al*, 2007). In order to allow the model parameters to vary in time only slightly, let σ = 0.00001.

Observation equation

The monitoring samples S_t come from successive months (t = 1, 2, …). On the basis of each sample S_t the model parameter estimates $\hat{\beta}_t^m$ are found. While the parameters β_t determine the state of a process, those estimates are used as a measurement. The relationship between them is described by the observation equation, which is supposed to have the following form:

$$\hat{\beta}_t^m = \beta_t + r_t ,$$

where var(r_t) = R_t for all t. Because there is neither assumption on the scoring model specification nor assumption on the estimation method, the estimator features are unknown. However, it seems safe to assume that the estimator is unbiased and follows an asymptotic multivariate normal distribution:

$$\hat{\beta}_t^m \sim \mathrm{N}(\beta_t, R_t).$$

Contrary to Whittaker *et al* (2007), there is no reason to presuppose any specific form of the estimator covariance matrix R_t. In particular it would be unjustified to assume, like Whittaker *et al* (2007), who use the maximum likelihood method, that the matrix R_t is an inverse of the Fisher information matrix. Such an assumption would be unjustified because it is not known whether the estimator is the most efficient one. Therefore, an estimate of the matrix R_t is derived from the parametric bootstrap.

In order to perform the bootstrap, a new sample B_t is chosen from the original one S_t, using proportional sampling with replacement (i.e. with repetition allowed). As a result the new sample is equal in size to the original one and the odds are the same. On the basis of the sample B_t the model parameters are estimated and then the obtained estimates are collected. The sampling and the estimation are repeated, for example, 100 times. The collected parameter estimates are used to compute the covariance matrix which constitutes the bootstrap estimate of the matrix R_t.

Updated scorecard

The updated scorecard is the output of the Kalman filter. Its parameter estimates $\hat{\beta}_t$ are found using the Kalman filter on the n-dimensional state space, where n is a number of the model parameters. Those estimates are the actual ones, while the estimates $\hat{\beta}_t^m$, which are obtained on the basis of a monitoring sample, are treated as only a noisy measurement.

In order to calculate the Kalman filter estimates, the time and measurement update equations are used (see Welch and Bishop, 2006). Firstly, using the time update equations, the *a priori* estimates $\hat{\beta}_t^-$ are determined and the *a priori* error covariance matrix P_t^- is computed:

$$\hat{\beta}_t^- = \hat{\beta}_{t-1},$$
$$P_t^- = P_{t-1} + Q.$$

In this case the current *a priori* estimates are equal to the previous *a posteriori* estimates. Secondly, the Kalman gain K_t is calculated according to the following formula:

$$K_t = P_t^-(P_t^- + R_t)^{-1}.$$

Finally, using the measurement update equations, the *a posteriori* estimates $\hat{\beta}_t$ are found and the *a posteriori* error covariance matrix P_t is computed:

$$\hat{\beta}_t = \hat{\beta}_t^- + K_t(\hat{\beta}_t^m - \hat{\beta}_t^-),$$
$$P_t = (I - K_t)P_t^-.$$

The parameter estimates $\hat{\beta}_t$ of the updated scorecard are determined on the basis of the *a priori* estimates $\hat{\beta}_t^-$, the Kalman gain K_t and the estimates $\hat{\beta}_t^m$ from the monitoring sample.

As far as initial values are concerned, it is assumed that the *a priori* estimates $\hat{\beta}_1^-$ are equal to the parameter estimates $\hat{\beta}_0^b$ of the baseline model:

$$\hat{\beta}_1^- = \hat{\beta}_0 = \hat{\beta}_0^b.$$

According to Whittaker *et al* (2007), the initial error covariance matrix P_0 should be such that the baseline model parameter estimates have a relatively weak influence on the estimates $\hat{\beta}_1$. Therefore, it is supposed that $P_0 = 10000I$. As a consequence, the estimates $\hat{\beta}_t$ are affected more by the estimates $\hat{\beta}_t^m$ from the monitoring sample than by the parameter estimates of the baseline model. The updated scorecard is determined on the basis of both the current monitoring sample and the previous ones but it depends more on the former than on the latter.

Linear model

It is common practice that once a scorecard has been developed, an additional linear model is estimated, in order to find a relationship between the score and the natural logarithm of the odds for that score (see Mays, 2004, p 71). That linear relationship is often used to scale the scorecard, i.e. to change the model parameters so that there is a required dependency between the score and the odds or the probability of the modelled phenomenon (see Siddiqi, 2005, p 113). It is especially useful when an institution (a bank or a credit bureau) has several scorecards and wants them to be consistent in terms of scale.

In order to estimate an additional linear model, the whole score range is divided into *m* equal-length ranges. The model is developed on the basis of the data on the mid-points and the log odds of those ranges (see Mays, 2004, p 71). Some customers with the lowest and highest scores can be treated as outliers and thus excluded from the model estimation.

In this research, additional linear models are built for both the baseline and updated scorecards. For the baseline scorecard, the following model is assumed:

$$\ln(\hat{o}_i) = \hat{a}_0^b + \hat{b}_0^b \cdot s_i^b ,$$

where s_i^b is a score coming from that scorecard. The baseline linear model is estimated on the basis of the sample S_0. The above relationship is used to determine the customer's credit propensity according to the baseline scorecard. Using the point estimation, one could predict that a customer *i*, whose baseline score equals s_i^b, is willing to apply for new loans at the level corresponding to the odds $\hat{o}_i$. Using the interval estimation, one could obtain the 90% confidence interval $(l_i^l, l_i^u) = \left(\ln(\hat{o}_i) - t_* S_i, \ln(\hat{o}_i) + t_* S_i\right)$ of the log odds, such that:

$$\mathrm{P}\left\{\ln(\hat{o}_i) - t_* S_i < \ln(o_i) < \ln(\hat{o}_i) + t_* S_i\right\} = 0.9 ,$$

where t_* is the appropriate value of the Student's distribution with *m*–2 degrees of freedom and S_i is the *ex ante* forecast error (see Greene, 2000, p 307).

As far as the updated scorecard is concerned, the linear model, which is estimated on the basis of the monitoring sample S_t, takes the following form:

$$\ln(\hat{o}_i) = \hat{a}_t + \hat{b}_t \cdot s_i,$$

where s_i is a score that comes from the mentioned scorecard. The above relationship enables the customer's credit propensity to be determined according to the updated scorecard. Using the point estimation, it could be predicted that a customer i, whose updated score is equal to s_i, is willing to apply for new loans at the level corresponding to the odds $\hat{o}_i$.

Performance monitoring

Each customer belonging to the monitoring sample is scored using both the baseline scorecard and the updated one. Both scores are calculated on the basis of the customer's data for the same moment. Then the log odds are estimated using the baseline linear model and the linear model for the updated scorecard, respectively. Those odds are treated as measures of the customer's credit propensity according to the baseline and updated scorecards. Provided that the baseline scorecard is still up-to-date, the odds should not differ too much from each other. In particular, the log odds estimate, which is obtained on the basis of the updated score, should in principle lie within the 90% confidence interval determined using the baseline score of the customer for the same moment. If the estimate does not fit within the interval, the baseline and updated scorecards differ considerably in their assessment of the customer's credit propensity level. If there are numerous cases like that, one can conclude that the baseline scorecard is not up-to-date in terms of assigning the odds and probability of applying for a new loan. Therefore, the percentage of customers, for whom the above-mentioned condition is not fulfilled, is analysed for each monitoring sample.

Simultaneously, the scorecard performance measures, the Gini coefficient and the KS statistic, are tracked in order to verify the discriminatory power of the baseline scorecard over successive months. In propensity scoring the Gini coefficient is a measure of ability to rank customers according to their credit propensity while the KS statistic measures ability to separate the willing from the unwilling. However, even if the ranking and separation statistics remain unchanged, the relationship between the score and the log odds can change

considerably (see Mays, 2004, p 116). It can mean that the credit propensity level is systematically under- or overestimated. Therefore, the propensity scorecard monitoring should include an analysis of the mentioned relationship. Such an analysis usually consists of comparing the actual (empirical) odds with their estimates obtained using the baseline linear model. The baseline scorecard performance is assessed using one monitoring sample each time. Thus, a single untypical sample can lead to a negative monitoring result and redevelopment of the model.

However, in this paper that approach is replaced with tracking the percentage of customers whose updated odds do not lie within the 90% confidence intervals determined using their baseline scores. Thus, the baseline scorecard performance is assessed using not only the current monitoring sample, but – through the Kalman filter – all previous ones as well.

Empirical results

Data

The presented example is based on the credit bureau data consisting of twelve samples: a baseline sample and eleven monitoring ones. Each sample has a different observation point. Those observation points are derived from twelve successive months. The outcome period always equals four months here. In each sample there are customer's characteristics (variables) as of the observation point and a customer's status (willing or unwilling) as of the outcome point.

The baseline scorecard is developed using a random sample consisting of 6309 customers (including 1229 willing ones) whose data are collected in the BIK database. The observation point is the 1st of September 2005 and the outcome point is the 1st of January 2006 (four months later). The sample is also used to develop the baseline linear model. The monitoring samples, which are used to calculate estimates serving as measurements in the Kalman filter, come from eleven successive months (t = 1, 2, …, 11). Each monitoring sample consists of over six thousand customers randomly selected from the database. In the consecutive samples there are the following observation points: the 1st of October 2005, the 1st of November 2005, …, the 1st of August 2006. Because a four-month outcome period is assumed, the respective outcome points are: the 1st of February 2006, the 1st of March 2006, …, the 1st of

December 2006. Both in the baseline sample and in the monitoring ones the odds are similar and equal ca 4.

Model parameters

The developed model has 29 parameters (nine characteristics in the form of binned variables). Since this is a propensity scorecard based on credit bureau data, the variables describe the customer's credit history and credit activity (especially within the last year). There are such characteristics as: number of credit inquiries within the last 12 months (0, 1, 2 or 3 and more), number of loans granted within the last 12 months (0, 1 or 2 and more), number of past loans (0, 1, 2-3 or 4 and more), time since last credit inquiry (below 6 months or 6 months and above, or no inquires), and number of different products applied for within the last 12 months (1 or 2 and more, or no inquires). The model parameters are estimated using commercial software dedicated to scorecard development. The estimation method is not mentioned in the software documentation and thus the estimator features remain unknown.

Scorecard monitoring

The baseline scorecard is monitored using the Kalman-filter-based technique described in this paper. In the beginning (for $t = 0$) there is no updated model. However, for $t = 1$ the baseline scorecard is treated as if it were an updated one from the preceding moment (in order to determine the initial *a priori* estimates of the Kalman filter). Therefore, in the beginning the updated model can be assumed to be the same as the baseline one, and the tracked percentage of customers is equal to zero. The next updated scorecard is the first output (*a posteriori* estimates) of the Kalman filter. The model parameter estimates, which are obtained on the basis of the first monitoring sample, constitute the measurement used to produce that output. The parameter estimates of the updated scorecard serve as the *a priori* estimates, which are then transformed into the *a posteriori* ones using estimates from the second monitoring sample (the next measurement). The *a posteriori* estimates constitute the second updated scorecard. They serve then as the *a priori* estimates used (together with the new measurement) to estimate parameters of the third updated scorecard, and so on. As a result, there is a sequence of updated scorecards. As an example, measurements and updated attribute scores of the selected characteristic (number of past loans) are presented in Figure 1.

For each updated scorecard an additional linear model is built (based on the monitoring sample).

The baseline scorecard is observed over a year. For each month customers from the monitoring sample are scored using that model. Then the Gini coefficient and the KS statistic of the baseline scorecard are computed. The updated scorecard and linear model are used to calculate the updated odds for each customer. It is checked whether those odds fit within the 90% confidence interval determined using the customer's baseline score and linear model. The percentage of customers, whose updated odds do not fit within the intervals, is computed.

Scorecard performance

All the tracked measures are presented in Table 1 and illustrated in Figures 2, 3 and 4. In the beginning (for $t = 0$) the Gini coefficient and the KS statistic of the baseline scorecard equal ca 0.43 and 0.33, respectively. However, in the monitoring period they are slightly lower. Although they remain relatively stable over time, there is some evidence that the baseline scorecard has deteriorated. The percentage of customers, whose updated odds do not lie within the 90% confidence intervals determined using their baseline scores, increases generally over the successive months. The observed tendency is clear: the credit propensity level is either under- or overestimated for the increasing percentage of customers. After eight months of the model monitoring, for $t = 8$, the tracked percentage exceeds 20%, which means that more than one in five customers has an updated odds lying beyond the interval. It seems that one could expect this to increase further and, as a consequence, further degradation of the baseline scorecard in the subsequent months could be also expected.

The obtained results could be interpreted in the following way. Since the discriminatory power measures are reasonably stable, the scorecard retains its ability to separate the willing from the unwilling as well as to rank customers according to their willingness to apply for new loans. However, there is an increase in the percentage of customers whose updated odds do not fit within the determined intervals. Thus, the successive updated scorecards differ more and more from the baseline one in their assessment of the customer's credit propensity level. This could be interpreted that in the consecutive months the model becomes less and less up-to-date in terms of assigning the odds and probability of applying for a new loan.

Conclusions

The presented example demonstrates that a scorecard may become less up-to-date, although the commonly used performance measures such as the Gini coefficient or the KS statistic do not change considerably. It is up to the decision makers as to what the maximum value of the analysed percentage that can still be accepted is (probably 10% would be a good idea in the case that the 90% confidence intervals are used). Once such a value has been exceeded, the model has to be redeveloped (or a completely new model should be built). Using a degraded scorecard may result in wrong business decisions and thus is not recommended, especially in the case of a cut-off determined on the basis of a relationship between the score and the odds for that score.

Kalman filtering can help detect such scorecard failures in the monitoring process. One of the main advantages of that technique seems to lie in using not only the current monitoring sample but – through the Kalman filter – all previous ones as well. In effect, possible local disturbance should have a limited influence on the monitoring results and thus on the decisions based on them (poor monitoring results can indicate that a new model should be built). Another advantage of the approach, which is presented in this paper, lies in the lack of assumptions concerning both model specification and estimation method (often the case in practical applications based on commercial software). Thus the demonstrated technique seems useful as a monitoring tool for different scorecards, including, but not limited to, propensity ones.

A monitoring result, which indicates that the scorecard does not assign odds correctly, is of great importance for the scorecard user. However, it is sometimes more important to know whether the odds are systematically under- or overestimated. In credit scoring overestimated odds mean underestimated credit risk. In such a situation, score-based credit decisions may result in an unexpected decrease in the bank portfolio quality. In the reverse situation an excessive number of applicants are rejected, which reduces the bank profit. In propensity scoring, the underestimated odds seem to be a more serious problem than the overestimated ones, because among customers selected for a marketing campaign there are less willing ones than expected. Thus, the response rate may be lower than assumed, which has a negative influence on the campaign efficiency. Therefore, further modifications of the presented technique could more specifically distinguish between under- and overestimation of the odds.

Month	Percentage of customers	Gini coefficient	KS statistic
Sep-05	0.0%	0.431	0.327
Oct-05	7.2%	0.359	0.263
Nov-05	3.9%	0.363	0.289
Dec-05	11.8%	0.350	0.275
Jan-06	4.7%	0.406	0.309
Feb-06	8.3%	0.370	0.278
Mar-06	14.3%	0.349	0.260
Apr-06	9.5%	0.355	0.265
May-06	22.1%	0.370	0.276
Jun-06	13.3%	0.375	0.276
Jul-06	16.5%	0.376	0.282
Aug-06	18.4%	0.406	0.322

Table 1. The monitoring results

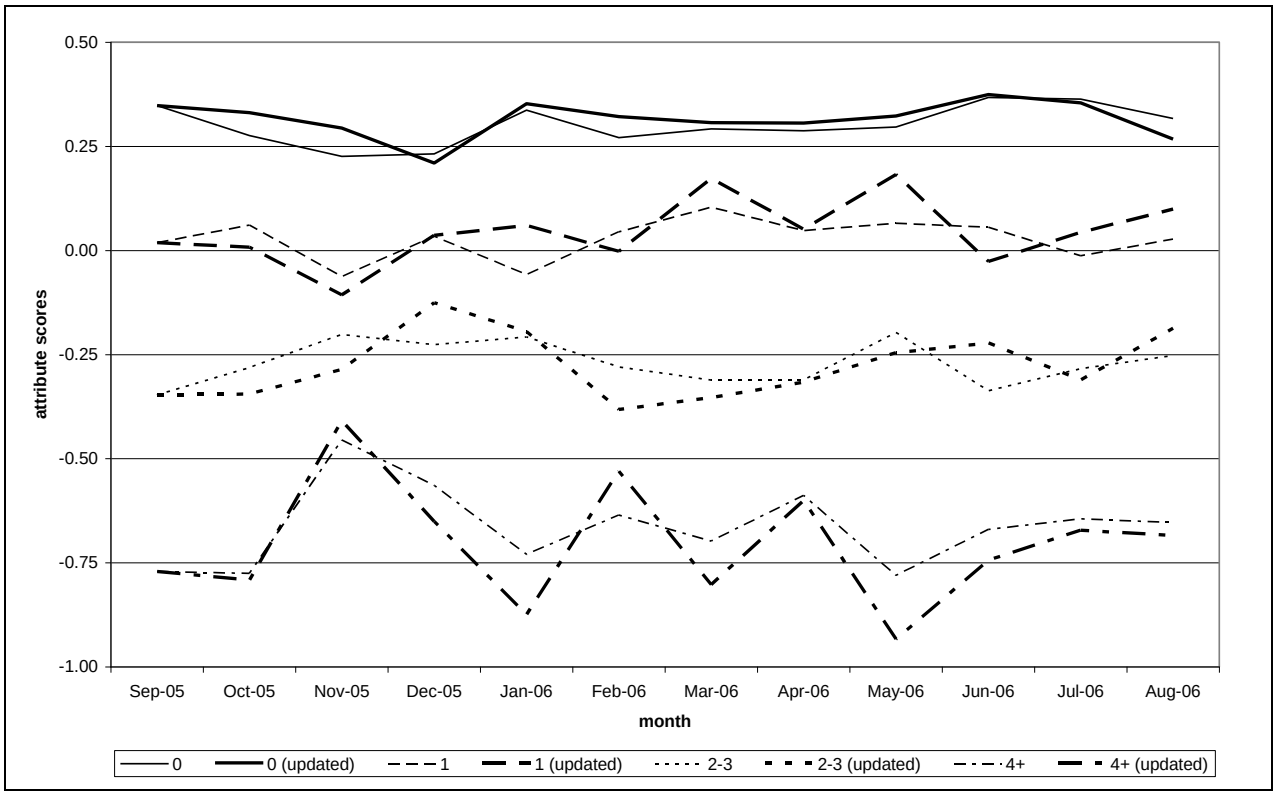

Figure 1. Attribute scores of the selected characteristic (number of past loans = 0, 1, 2-3 or 4 and more): estimated on the basis of successive samples and updated using the Kalman filter

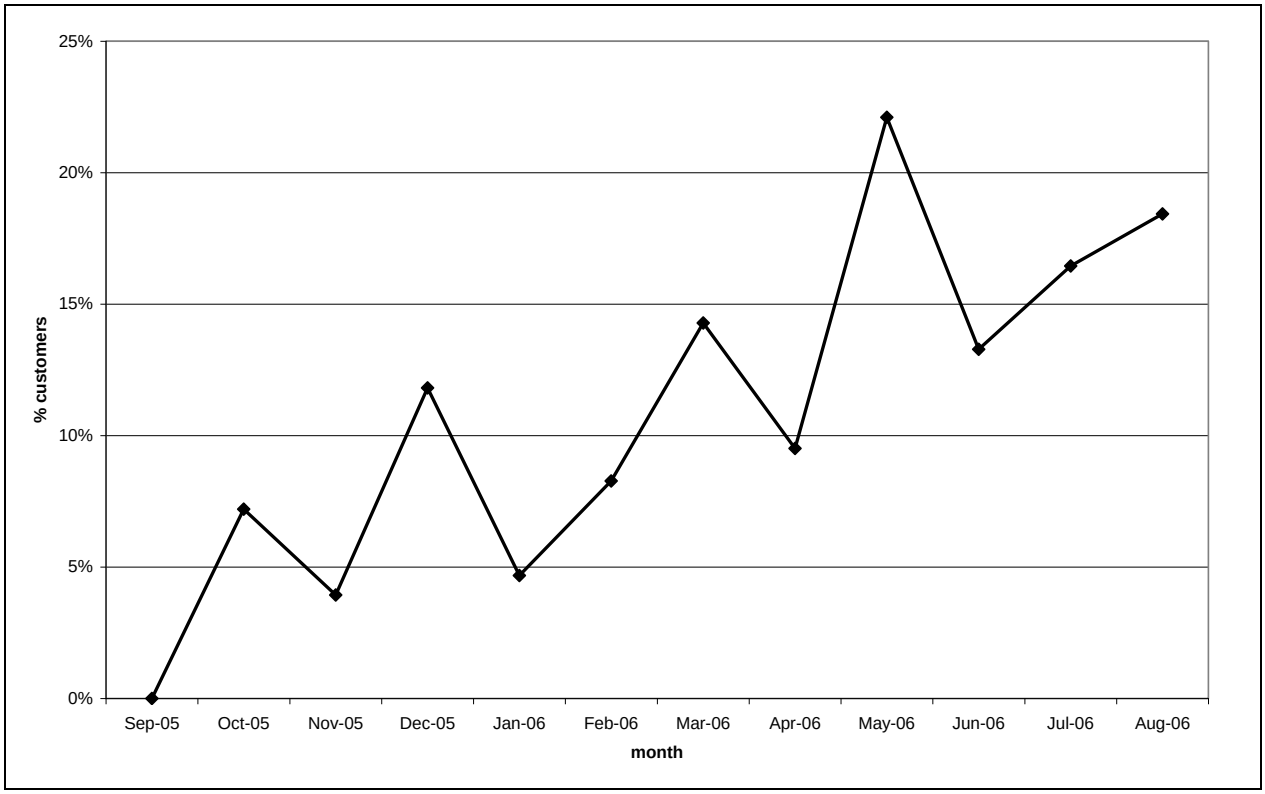

Figure 2. The percentage of customers whose updated odds do not lie within the intervals

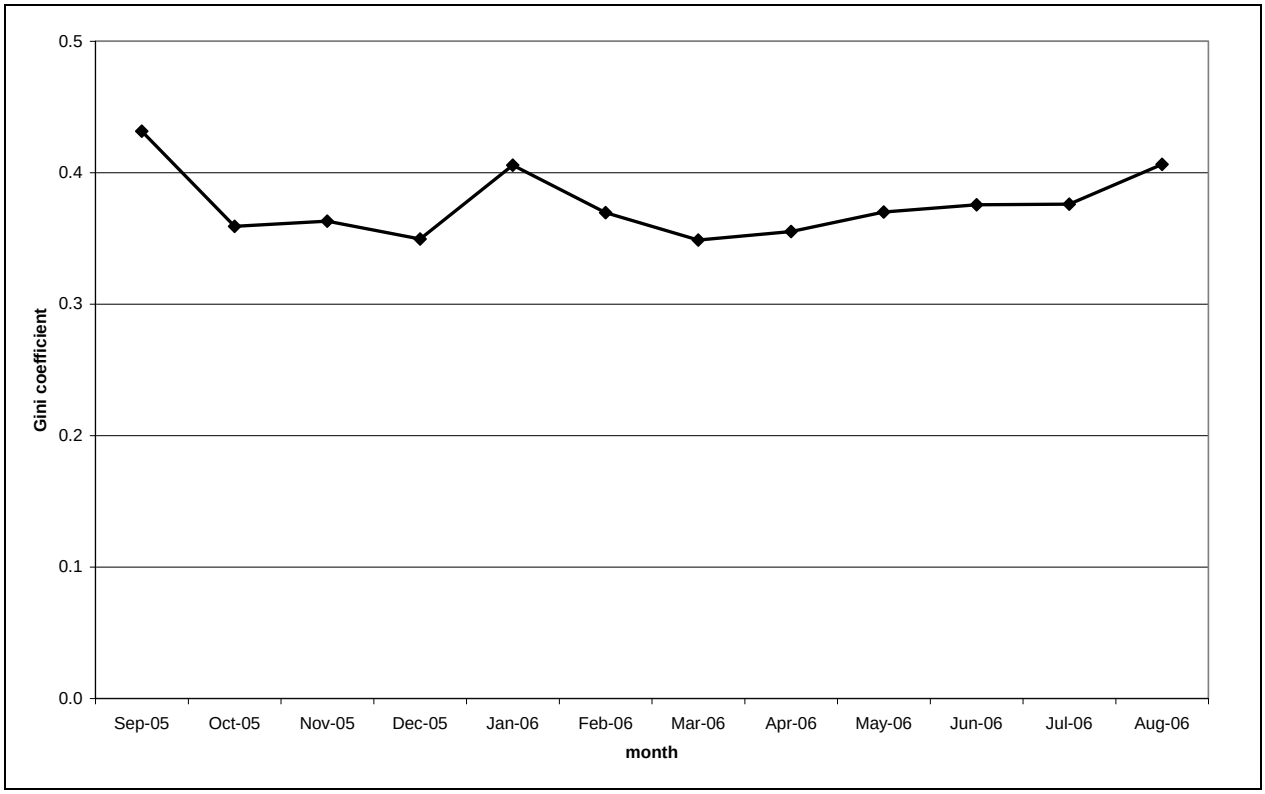

Figure 3. The Gini coefficient of the baseline scorecard

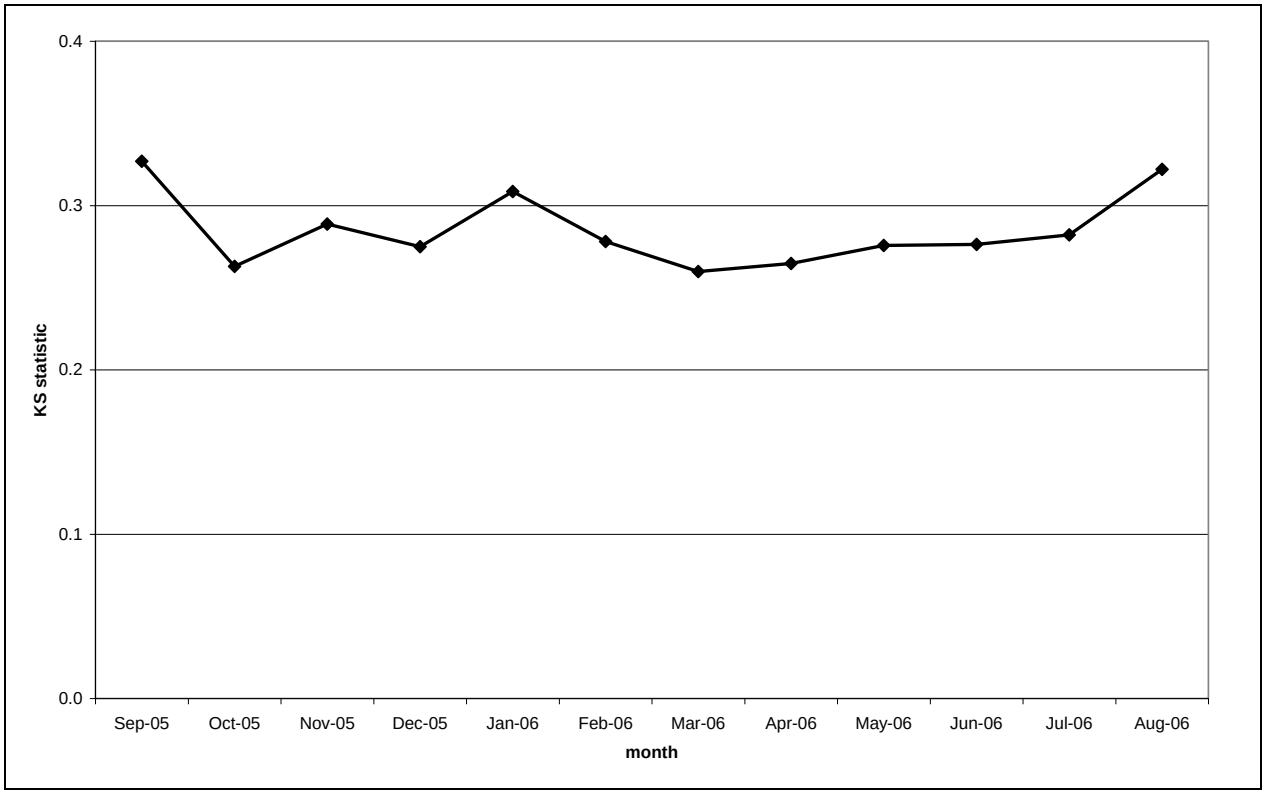

Figure 4. The KS statistic of the baseline scorecard

References

Anderson R (2007). The Credit Scoring Toolkit. Oxford University Press: New York.

Greene WH (2000). Econometric Analysis. Prentice Hall: Upper Saddle River.

Harvey AC (1990). Forecasting, structural time series models and the Kalman filter. Cambridge University Press: Cambridge.

Kalman RE (1960). A New Approach to Linear Filtering and Prediction Problems. Transactions of the ASME – Journal of Basic Engineering 82 (Series D): 35-45.

Lucas A (2004). Updating scorecards: removing the mystique. In: Thomas LC, Edelman DB and Crook JN (eds). Readings in Credit Scoring: Foundations, Developments, and Aims. Oxford University Press: New York, pp 93-109.

Mays E (2004). Credit Scoring for Risk Managers. The Handbook for Lenders. Thomson South-Western: Mason, Ohio.

Schiffman R (2001). Evaluating and Monitoring Your Model. In: Mays E (ed). Handbook of Credit Scoring. Glenlake Publishing Company, Ltd.: Chicago, pp 285-300.

Siddiqi N (2005). Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. Wiley: New York.

Thomas LC, Edelman DB and Crook JN (2002). Credit Scoring and Its Applications. SIAM: Philadelphia.

Thomas LC (2009). Consumer Credit Models: Pricing, Profit, and Portfolios. Oxford University Press: New York.

Van Gestel T and Baesens B (2009). Credit Risk Management. Basic concepts: financial risk components, rating analysis, models, economic and regulatory capital. Oxford University Press: New York.

Welch G and Bishop G (2006). An Introduction to the Kalman Filter. Tech. Report TR 95-041, Department of Computer Science at the University of North Carolina at Chapel Hill. http://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf, accessed 2 June 2008.

Wells C (1996). The Kalman Filter in Finance. Kluwer Academic Publishers: Dordrecht.

Whittaker J, Whitehead C and Somers M (2007). A dynamic scorecard for monitoring baseline performance with application to tracking a mortgage portfolio. J Opl Res Soc 58: 911-921.

Figure 1. Attribute scores of the selected characteristic (number of past loans = 0, 1, 2-3 or 4 and more): estimated on the basis of successive samples and updated using the Kalman filter

Figure 2. The percentage of customers whose updated odds do not lie within the intervals

Figure 3. The Gini coefficient of the baseline scorecard

Figure 4. The KS statistic of the baseline scorecard

Table 1. The monitoring results