DATA MINING BASED MODEL AGGREGATION

SZŰCS, IMRE

Keywords: model aggregation, consistent future, data mining, CRM, Basel II.

CONCLUSIONS

Applying modelling techniques for getting acquainted with customer behaviour, predicting the customers' next step is neccessary to keep in competition, by decreasing the capital requirement (Basel II - IRB) or making the portfolio more profitable. According to the easily implementable modelling techniques, data mining solutions widespread in practice. Using these models with no conditions can lead into inconsistent future on portfolio change. Consequence of this situation, contradictory predictions and conclusions come into existence. Recognizing and conscious handling of inconsistent predictions is an important task for experts working on different scene of the knowledge based economy and society. By realizing and solving the problem of inconsistency in modelling processes, the competitive advantage can be increased and strategic decisions can be supported by consistent predictions.

ABSTRACT

In recent years, data mining based modeling techniques widespread in several sector of the economy. As competition becomes more and more close now, one of the main competitive edge for market prticipants is to utilize sophisticated modelling techniques to support CRM and marketing activity. Furthermore, due to the New Basel Capital Accord for financial organizations, there is a must to develop as sophisticated models as possible to avoid the growth of capital adequacy. Thanks to these effects, several models were developed to predict customers' behaviour based on their known attributes. In this paper, I show that transaction level models perform better then customer level models, but the interpretation of transaction level models on customer level might drive to inconsistency. To resolve this problem, I analyse different methods for aggregation such as expert method, linear regression and neural network. Finally, I evaluate the efficiency of different methodologies.

INTRODUCTION

Using widespread data mining techniques in every day practice for predicting different customer flavour, several models models came alive. Response and churn models supports marketing activity, risk models predicts the requirement of the new basel capital accord. The results of these models can be used

1. as an order of customers priority. According to this priority the target of the offer can be optimized;
2. for calculating the profitability and the risk parameters.

Using these models without any compromise may lead to inconsistent future.

1st type of inconsistency. A typical example when we develope different models for customers with different products for predicting the possibility of new product purchase. Due to problem of the missing values, developing only one model for all customers is not efficient. Having models for each product causes the equality of the estimated parameters with low possibility.

2nd type of inconsistency. We face the same problem when estimating the parameters (PD – Probabilty of Default, LGD – Loss Given Defaultm, EAD – Exposure At Default) required by the new basel capital accord. In case of using all available information related to other products of the customer (not only the the examined product information) lead to the same problem mentioned above.

It is also an important issue that we cannot take into account the reaction of the competitors and the possible changing of the economical enviroment. According to this, using the models for calculation results inconsitent future. It predicts such change in the portfolio that is irreal in the current economical enz competitive circumstances. In this study we show an examlpe on the topic of the 1st type of inconsistency, and try to find some aggregating techniques as a solution for the inconsistency problem.

MATERIAL

The predicted event was the product purchase, but the results can be implememted for PD modelling as well, when the event is the customer default. The modelling basis were customers own product_1 or product_2, which products are not the same as the offered one.

The models that predict the purchase possility (p1 and p2) are developed based on the following data:

Table 1

Descriptive information on base models

	Model_1	**Model_2**
Objects	Customers	Customers
Attributes	Demographical and product_1 related data	Demographical and product_2 related data
Target variables	Purchase event	Purchase event
Good ratio	33,33%	33,33%
Expected value of purchase in validation sample	33,33%	33,33%
Misclassification error at cutting point p=0.5	22,60%	21,25%
Average squared error	0,1544	0,1562

INCONSISTENCY

Both models can be used for customers who own product_1 and product_2. In 46.38% of the sample, the value of the target variable is 1. To measure the model performance on this sample, several index numbers can be calculated.

One of the simpliest, however, most misleading solution is to compare the expected target value based on the estimated parameters and the frequency of the target variable of the sample. Classifying objects with higher probability then 0,5 to the group, in which the target variable value is 1, is a widespread method in practice to measure the performance of the models. The misclassification rate column shows this type of error. One of the simpliest calculatable type of error of the models is the average squred error, can be found in column

ASE. The presentation of the separating capability of a model, in case of binary target variable, is the ROC (Receiver Operating Characteristic Curve). To give only one number which characterize the model, the area under the ROC (A) can be used. Another method for visualizing the separating capability of a model is the CAP (Cumulative Accuracy Profile) curve. AR is to numerically typify the model. There is a linear connection between A and AR: AR = 2(A-0.5). (Engelmann, Hayden and Tasche)

Table 2

Input variables used for modeling

Variable	Type	Status	Dimension	Minimum	Maximum
Customer ID	id	1,2	identifier		
Customer segment	nominal	1,2	category		
Age	continous	1,2	year	24	75
Income	continous	1,2	HUF	0	1,500,000
Sum of bank card transaction	continous	1,2	HUF	0	3,000,000
Cash flow in branch bank	continous	1	HUF	0	30,000,000
Account debits	continous	1	HUF	0	100,000,000
Account credit	continous	1	HUF	0	200,000,000
Customer contact time	continous	1,2	month	0	70
VIP flag	ordinal	1,2	category	0	5
Profitability	continous	1,2	HUF	-500 000	5,000,000
Number of products owned by customer	continous	1,2	pieces	0	9
Loans amount	continous	1,2	HUF	0	40,000,000
Sum total of deposits	continous	1,2	HUF	0	50,000,000
Mortgage amount	continous	1,2	HUF	0	50,000,000
Personal loans amount	continous	1,2	HUF	0	5,000,000
Credit card loan amount	continous	2	HUF	0	1,000,000
Sum of total time deposit	continous	1,2	HUF	0	50,000,000
Overdraft loan amount	continous	1	HUF	0	5,000,000
Account balance	continous	1	HUF	0	100,000,000
Security value	continous	1,2	HUF	0	5,000,000
Credit card type	nominal	2	category	A	E
Credit balance utilization	continous	1	Percent	0	1
Maximum of credit limit utilization	continous	2	HUF	0	1,000,000
Sum of credit card transactions	continous	2	pieces	0	50

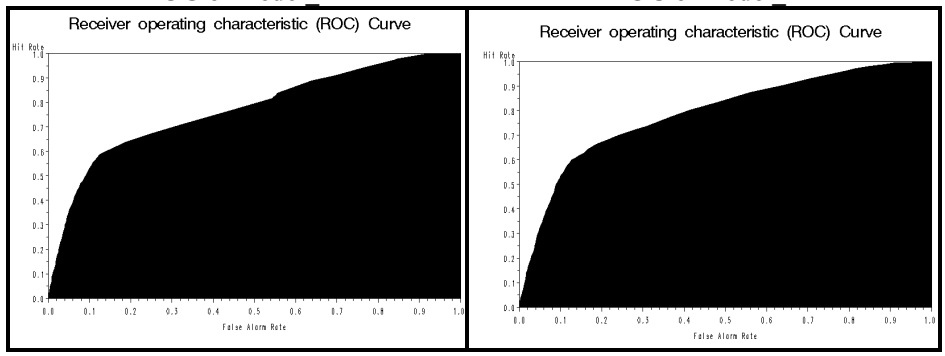
Table 3

Model_1 and model_2 comparision (own calculation)

	Expected value of product purchase	Misclassification error (at p=0,5)	Average squred error ASE	Area under ROC	Accuracy Ratio
Model_1	0.4269	26.73	0.1941	0.7718	0.5437
Model_2	0.3741	25.25	0.1951	0.4910	0.5821

Figure 1 **Figure 2**

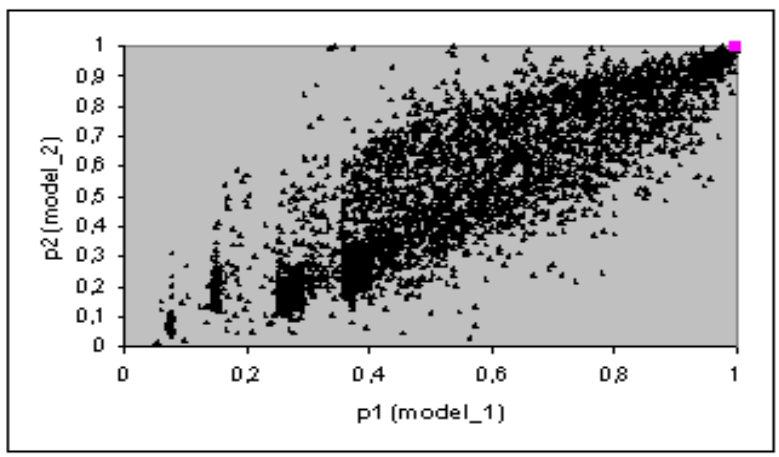
ROC of model_1 **ROC of model_2**



Considering any feature the base models perform worth comparing to the development sample (Table 3). Regarding the customers, one possible reason is that there is no full picture about them in the base model creation phase. Comparing the models to each other its obvious that base models predict different result for the target variable (Figure 3).

Figure 3

Result of model_1 and model_2 on the common sample



Source: own calculation: 8500 points

Table 4

Correlation between the predicted parameters of model_1 and model_2

Correlation	Correlation coefficient
Pearson	0.8977
Spearman	0.8898

Based on the results above it can be seen that the correlation between the predicted possibility is relatively high, but there are differences between the numeri-

cally values predicted for custumoers. High correlation is a requirement because all the models should represent the reality. Differences can be justifiable with the fact, that in base model creation phase the whole customer information is not available. The problem turns into significant as long as the predicted parameters are used in further calculations, such as:

- Expected Loss determination:

$$EL = PD * EAD * LGD, \quad (1)$$

where PD – Probability of Default, EAD – Exposure at Default, LGD – Loss Given Default.

If the difference is significant, it causes incorrect level capital.

- Customer Value determination:

if the customer value determination uses the response and churn models as input parameters. In this situation, the cumulative difference makes the customer value impossible to priorize the customers.

METHODS – SOLVING INCONSISTENCY

There are several methods for solving inconsistency

1. Common model development
2. Model development for each possible product combination
3. Model aggregation
   a. Expert method
   b. Linear regression
   c. Artifical neural network
   d. Component-based Object Comparison for Objectivity (COCO)

Common model development

As a roundabout process one common model can be developed which use each customer and each product in modelling the target variable. In this case we face the following problem. For customers who do not own a product, the columns related to this product will be filled with NULL values. Having discrete variables the missing values can be replaced with a MISSING category, but in case of continous variables the replacement technique makes them similar to customers with non missing values. This approach is not valid from business point of view.

Model development for each possible product combination

Theoreticaly there is a possibility to develop individual model for each possible product combination. In this way the prediction will use all available information related to each customer. In pracitce this approach results 75 models in case of a bank with 5 different products. The development and the management of such number of models need significant work, but unfortunately there is no practically smooth methodology for it. It is an additional problem that there is not enough good or bad customer in each product combination's clientele.

Model aggregation

Developing distinct models for each product, predicting the target product purchase the problem can be simplified. Both the number of models and the difficulty of managing the models decrease drastically (Table 5). Here as an additional problem the inconsitent future has to be solved by creating only one number to predict the product purchase. As aggregating methodds we examined the expert method, linear regression, artifical neural network and COCO.

Data used in model aggregation:

- Objects: customers own both product_1 and product_2
- Attributes: Predicted purchase probabilities (p1, p2) and errors of p1 and p2 (p1_ase, p2_ase).
- Target variable: product purchase event.

Table 5

Number of models by different approaches

Number of products N	Number of models in case of each product combination $N(2^{N-1}-1)$	Number of models in case of aggregation N(N-1)
2	2	2
3	9	6
4	28	12
5	75	20
6	186	30
7	441	42

Table 6

Input variables for aggregated model development

Variable	Type	Status	Dimension	Minimum	Maximum
Customer ID	id	---	id		
p1	continous	derived	probability	0	1
p2	continous	derived	probability	0	1
p1_ase	continous	derived	error	0	1
p2_ase	continous	derived	error	0	1
Target product purchase	discrete	fact	yes / no	0	1

MODEL AGGREGATION – EXPERT METHOD

Using the expert method we try to involve the knowledge: rather use the „better" model. To reach it we calculate the final predicted parameter as the inverse ratio to error wighted average of the base models predicted parameters. It is easy to understand in business point of view but obviously it is not an optimized solution. The main question regardind to this solution is how to measure hte error. In practice there are several method used widely:

- SSE - Sum of Squared Error:
- ASE – Average Squared Error = SSE / N
- MSE – Mean Squared Error = SSE / (N-P)
- RMSE – Root Mean Squared Error = $(MSE)^{½}$
- FPE – Final Prediction Error (Akiake) = SSE * (N+P) / [N * (N-P)],

where N is the number of objects and P is the number of the predicted weights.

Table 7

Error numbers of the base models on their own validation sample

	ASE	MSE	RMSE	FPE
Model_1	0.1544	0.1547	0.3933	0.1550
Model_2	0.1562	0.1570	0.3962	0.1578

Using te ASE for defining the badness of the models we can define the following equation for the final product purchase probability as an expert approach:

$$p = (ASE\_1 * p2 + ASE\_2 * p1) / (ASE\_1 + ASE\_2). \quad (2)$$

Table 8 shows the comparison of the 2 base models to the aggregated model on

the common sample (customers who own both product_1 and product_2). As it can be seen the aggregated model works better than the base models on the common sample, but does not reach the ASE they provided on their own validation sample. Our main goal is to minimize this error value by aggregating techniques.

Table 8

Model comparison based on ASE

	Model_1	Model_2	Aggregated model
ASE	0.1940	0.1951	0.190923

To take into account not only the fact which model is the better, but for which customers which model is the better, it is necessary to segment the customer base. For this reason the customer base was segmented into 18 group by the variables p1 and p2. As a segmentation technique the K-mean clustering algorithm was used, where the distance metric was the Euclidean distance. Table 9 shows a few example on the types of segments can be found. In this way it is possible to indentify p1-p2 groups where one of the base models perform better than the other.

Table 9

Segments by p1 and p1

Segment	Average p1	Average p2	Average event
2	0.8546	0.7485	0.8957
5	0.697	0.7012	0.8233
14	0.4914	0.8448	0.8364
16	0.8241	0.8761	0.8638

Using equation 2 by segments a more sophisticated aggregated model can be made. For this model the ASE = 0,190920, better than the simple expert model only in the 6th decimal place.

MODEL AGGREGATION – LINEAR REGRESSION

In a problem like this, where the input variables are continous and the target variable can be handled as continous as well, the usage of linear regression seems to be trivial. As a part of the study linear regression was used to determine the coefficients, but better perform model was not found than the expert method was.

MODEL AGGREGATION – ARTIFICIAL NEURAL NETWORK

Applied artificial neural network:

- Multilayer Perceptron
- Activation function: tangens hiperbolicus
- Combination function: linear
- Hidden layers: 1-2
- Training – validation sample: 70%-30%

To make it easier the learning phase, the average squared error of p1 and p2 was modelled on customer level by memory based reasoning technique. In MBR the distance metric was the Euclidean distance, and the 50 nearest neighbours were take into account. During the neural network modeling the following effects were examined:

1. Using the predicted ASEs as an input parameter
2. Changing the number of hidden layers and the number of neurons.

In the first proceeding a two hidden layered network was used with 5 and 4 neurons.

Table 10

Comparing input parameters effects

Input parameters	Expected value of product purchase	Missclassification error at p=0,5	ASE	Area under ROC	Accuracy Ratio
P1, p2	0.4628	25.52	0.1778	0.8019	0.6038
p1, p2, p1_ase, p2_ase	0.4637	24.91	0.1760	0.8036	0.6072

Figure 4

ROC without ASE inputs

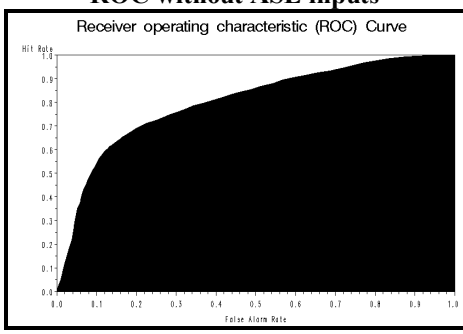
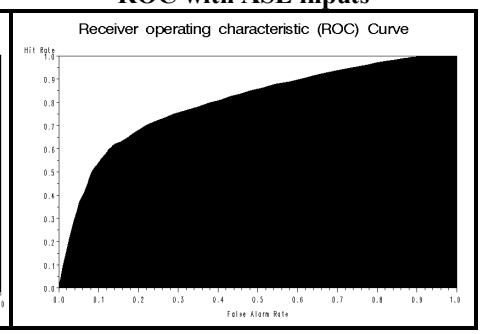


Figure 5

ROC with ASE inputs



Based on the results (Table 10) it can be established that it is worth to use the predicted ASE as input parameter. In this way the learning process of the neural network was forced by using parameters could be calculated inside the network as well. Thus means the predicted ASE usage help the learning only when the structure of the network is not complex enough for learning this pattern. It can be shown, that in more complex structured models the result of using predicted ASE has no added value for modelling. In the second proceeding the number of hidden layers and the number of neurons was changed. A few cases are shown in Table 11.

Network with 2 hidden layers perform better. But after a level of complexity it does not worth to increase the hidden layers number or the number of neurons from model performance point of view.

Table 11

Comparing different network structure

Number of neurons	Expected value of product purchase	Missclassification error at p=0,5	ASE
3	0.4659	25.89	0.1789
5	0.4659	25.58	0.1783
5 - 4	0.4628	25.52	0.1778
5 - 6	0.4637	24.93	0.1755

MODEL AGGREGATION – COMPONENT-BASED OBJECT COMPARISON FOR OBJECTIVITY (COCO)

As a part of the study the so-called COCO method was used for model aggregation as well (*Pitlik*). Using this method it is possible to develop the final model for prediction, and to quantify the importance of the input variables. The results of this method can be found in workpaper of *Pitlik, Szűcs, Pető, Pisartov and Orosz.*

RESULTS AND DISCUSSION

Each aggregating method solve the problem of inconsistence by giving only one value as the purchase probability. From performance point of view all of the examined method gave more accurate prediction than the base models on the two product owner customers' sample. The most precise prediction can be achieved by using artifical neural network for model aggregation. The performance of the neural network model can be increased by growing the network with increasing the number of hidden layers and neurons. But it is irreal to aim to get better prediction than the original base models could do.

REFERENCES

(1) http://miau.gau.hu/miau/92/nitra_full.doc – (2) http://miau.gau.hu/miau/91/bulletin_en.doc – (3) http://miau.gau.hu/miau/74/mtn2004_full.doc & http://miau.gau.hu /miau/74/cocomtn.xls – (4) http://miau.gau.hu/miau/81/iamo2005_en.doc & http://miau. gau.hu/miau/74/iamo_coco_poster.doc – (5) http://miau.gau.hu/miau/71/iamo_coco.doc – (6) http://miau.gau.hu/miau/91/gewisola2006_abstract.doc – (7) http://miau.gau.hu/ miau/73/gewisola_end.doc & http://miau.gau.hu/miau/73/gewisolafulltext.doc – (8) http://miau.gau.hu/miau/69/gilfull.doc – (9) http://miau.gau.hu/miau/85/gil26 _full.doc – (10) Pitlik, Szűcs, Pető, Pisartov and Orosz, 2006. Adatbányászati modellek aggregálása, May – (11) Basel Comittee on Banking Supervision, 2000. Supervisory Risk Assessment and early warning systems, December – (12) Basel Comittee on Banking Supervision, 2001. The internal rating-based approach Consultative document, January – (13) Sobehart J and S Keenan, 2001. Measuring default accurately Risk, pages S31-S33, March – (14) Jorge Sobehart, Sean Keenan and Roger Stein, 2000. Validation methodologies for default risk models Credit, pages 51-56, May – (15) Herman J. Bierens, 2005. Information criteria and model selection Pennsylvania State University, September – (16) Bernd Engelmann, Evelyn Hayden and Dirk Tasche, 2003 Testing rating accuracy www.risk.net, January