VIII Modelling Week, Madrid 2014

IMPACT OF MODEL ESTIMATION ON MODEL RISK

Master in Mathematical Engineering UCM

Instructors

Juan Tinguaro Rodríguez González (UCM)
Susana Fuster Lapesa (MS)
Berta Gómez Domínguez (MS)

Students

Antonio Aita
Marta González Marquina
Lorena Mellado Fernández
Cecilia Pérez Mazuela
Piotr Swierczynski

CONTENTS:

1. INTRODUCTION

Banks and other financial institutions rely heavily on models, which usually take the form of mathematical formulas. Most of the decisions made in the financial sector are supported by a prior quantitative analysis. Mathematical and statistical methods find applications in portfolio optimization, credit analysis, risk management and many other areas.

However, models are merely approximations of the reality and as such, inevitably lead to errors. Moreover, incautious implementation of mathematical models may become a source of even greater errors. This in turn may lead to huge losses made by the financial institutions. This observation justifies the introduction of the notion of model risk. According to Riccardo Rebonato (SEE, REFERENCE 1) model risk can be defined as "the risk of occurrence of a significant difference between the mark-to-model value of a complex and/or illiquid instrument, and the price at which the same instrument is revealed to have traded in the market". The notion of model risk is very complex and identifying sources of the risk can be very challenging since they may vary from model to model.

Nevertheless, main general sources of errors include:

- wrong model specification, e.g. oversimplifications in model building, neglecting a significant factor or variable,
- technical errors, e.g. wrong algorithm implementation or wrong algorithm choice,
- wrong model calibration and improper use of data.

One of the models most often used by banks is a scoring model. It is based on a formula that assigns points using known information to predict an unknown future outcome.

In this report we shall build such a model assessing the likelihood of a mortgage default by a bank customer.

We were provided with a dataset containing information about 100000 bank customers. This database covers a wide range of information such as customer's age, solvency or debit balance (SEE TABLE WITH DATA). Among them is the crucial from bank's perspective – information regarding mortgage default of each of bank's customers. Based on this database we build a model estimating the probability of the customer's default based on a variety of known parameters. Since the target variable is binary, logistic regression is a technique used for this purpose.

The outline of this report is as follows:

- We first perform statistical analysis and treatment of the dataset. We check whether all the variables are sufficiently informed by looking at the percentage of missing values. Missing vales are filled using several different approaches. We also find the extreme and unusual observations – the outliers – and present a way of treatment of such observations. Furthermore, for the sake of logistic regression's accuracy, we divide values of variables used in the regression into few subsets.

- Next, we move to the model building. We present a way of choosing the variables used of statistical model estimation. Then the selected variables are fitted to a logistic regression and the scoring model is obtained.

- In the next section we identify potential sources of errors in the built model. We also show the first attempt for the error quantification.

- Finally, we try to find a way of mitigating the error for each source identified before. Bootstrapping method is presented. Moreover, we propose a global risk measure quantifying impact of model estimation on the model risk.

2. MODEL CONSTRUCTION

2.1. Data Preprocessing

The first and one of the most important steps in constructing a good model is the data preprocessing. This process includes a number of data analysis techniques that improve the quality of the data to get more and better information.

The main points to be considered are the treatment of missing data, the treatment of outliers and the categorization of some of the variables.

MissingValues

In the following tables we can see all the possible variables for our model. Later, we are going to decide which of them are going to be part of our final model.

For the interval variables, we can see the minimum, the maximum, the mean, the standard deviation, the percentage of missing values, the skewness and de kurtosis:

Name	Min	Max	Mean	Std Dev.	Missing %	Skewness	Kurtosis
IDENTIFICADOR	1.1E7	6.13E7	3.54E7	1.44E7	0%	-0.074	-1.177
FECHA_APERTURA	13517	17895	15839	1221	0%	-0.289	-1.028
FINALIDAD	1	304	12.206	28.761	0%	5.7248	39.365
ANTIGUEDAD_CLIENTE	0	118.65	22.581	31.514	0%	1.2549	0.4469
ANTIGUEDAD_EMPLEO	0.0959	40.668	11.006	9.4235	17%	1.0872	0.4024
EDAD	1	82	37.399	10.708	16%	0.5391	-0.103
EDAD_HIJO_MENOR	1	43	12.342	8.9868	61%	0.6385	-0.315
ENDEUDAMIENTO	0	29.102	0.2306	0.9773	21%	22.333	579.61
LOAN_TO_VALUE	0.0065	68.634	0.6607	1.6131	5%	39.494	1656.8
SALDO_ACTIVO	-19.58	2.49E6	41508	135514	0%	10.453	156.97
SALDO_PASIVO	-14613	423408	8920.3	28480	0%	7.0723	68.792
SOLVENCIA	0	16.06	0.0921	0.5511	54%	26.565	764.23
TASA_ESFUERZO	0	64.866	0.3824	2.4326	17%	22.641	571.23
VAR_LIM_DIR_SISTEMA	-1	1	-56E-6	0.1301	43%	0.4537	21.535

There are some variables with a high percentage of missing values, for example *LDS* with 43% or *Solvency* with 54%. But the way to fill the missing values is not the same for both variables because of its character. We have deleted the variable *Age of the youngest children* due to its high percentage of missing values and the importance of that to our study.

For the class variables, we can see the number of categories they have and their percentage of missing values:

Name	Values	Missing %	Order
BANCO	8	0%	Ascending
CCAA	6	0%	Ascending
NIVEL_ESTUDIOS	6	0%	Ascending
PROFESION	5	0%	Ascending
ESTADO_CIVIL	7	0%	Ascending
NUMERO_HIJOS	6	16%	Ascending
NUMERO_TITULARES	7	0%	Ascending
IND_DEFAULT	2	0%	Descending
IND_INCIDEN_SISTEMA	3	0%	Ascending
IND_INCUMPLIMIENTOS	2	0%	Ascending
IND_NOMINA_DOMICILIADA	2	0%	Ascending
TIPO_CONT_LABORAL	3	0%	Ascending

We can observe that only the variable *Number of Children* has missing values. Because of the type of this variable, we fill its missing values in a different way that we have for the other ones.

Therefore, after several tests, we decide to approach the problem in three different ways:

- For some variables as *Solvency*, we replace the missing values by the 25 or 75 percentile, selecting the most conservative position for each case.

- Following this conservative point of view, for other variables as *Number of Children*, we replace the missing values for the worst case from the standpoint of business.

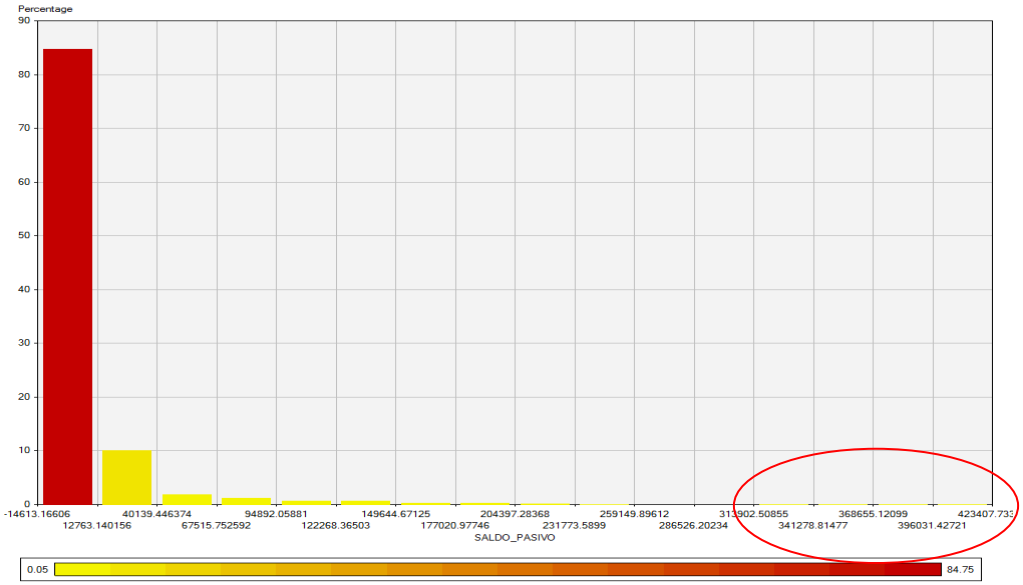
- For the variable *LDS*, a different decision is made. Due to the high percentage of missing values and its importance, we make a regression taking into account other variables related to it.

Outliers

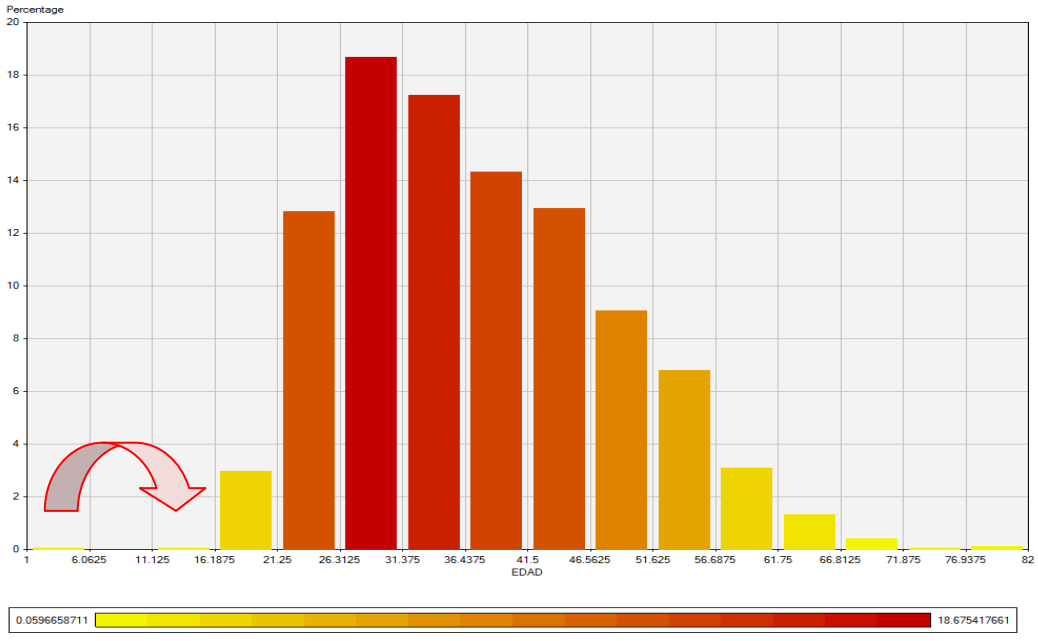
The next step is to decide the way we process the outliers. To have an idea of the situation with the outliers, we plot a histogram for each variable.

We approach the problem from two different points of view:

- If the outliers are extreme values, we delete all the observation that contains the outlier, taking into account that the percentage of these cases is small.
  For example we can clearly see in the histogram of the Debit Balance that there are some extreme values, and we have to delete these observations:



- If the outliers are out of the allowed range, we associate this case to a human error and we replace it by the nearest allowed value.
  For example we can clearly see in the histogram of the Age of the Customer that there are values less than 18, which is not possible:



Categorization

For some continuous variables we categorize them to check if it improves the relationship between the variable and the target. For example, the continuous variable

Age can be categorized in ranges in which the people's behavior is the same. The next table shows the categorization we choose:

AGE	
Class 1	[18, 30)
Class 2	[30,45)
Class 3	[45,65)
Class 4	≥ 65

We also made recategorizations for some categorical variables because we see that they have several categories all of them with the same sense of business. For example for the variable *Civil status* we join divorced and widowed people, and for the variable *Profession* we join civil servants and graduates.

2.2 Input Selection

Another step that we have to take into account is the input selection. Input selection can improve the performance and estimation of the classifier that predicts the likelihood of default.

In order to select variables, once they have been treated, we have to do a bivariant regression with each variable and compute the default index.

As the name suggests, in linear regression the relationship between the dependent and the independent variables is linear. This means that the predicted values could be greater than one and less than zero. On the other hand, this assumption is not made in logistic regression. Logistic regression estimates the probability of an event occurring (the predicted value lies within 0 and 1).

For this reason, in order to predict a binary response we have to use bivariant logistic regression.

After doing this, we should compare all of them by a performance measure and decide which ten variables are the best. In our case, we look at the Akaike information criterion (AIC) and theArea Under ROC curve (AUROC).

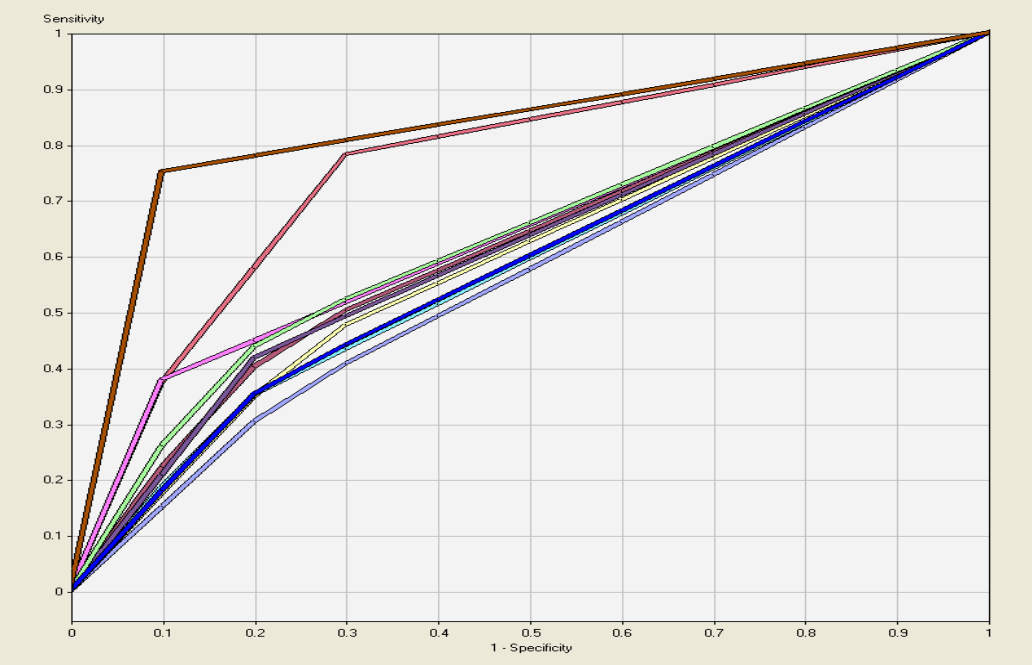
The following table shows these measures that are already sorted by AUROC (the bigger is the better):

VARIABLES	AIC	AUROC
BREACHES INDEX	23095,56	0,829
PURPOSE	27729,620	0,78
INCIDENCES IN THE SYSTEM INDEX	26639,30	0,766
SOLVENCY	30626,45	0,736
INDEBTEDNESS	30265,15	0,727
CUSTOMER SENIORITY	30104,520	0,721
DEBIT BALANCE	31366,90	0,651
LOAN TO VALUE	30548,73	0,644
EDUCATION LEVEL	30768,30	0,639
ASSET BALANCE	31365,32	0,635
PROFESSION	30968,28	0,627
CIVIL STATUS	30905,45	0,618
AGE	31385,36	0,616
VAR LIM DIR SYSTEM	31722,82	0,598
NUM CHILDREN	31429,50	0,582
NUM HOLDERS	31541,04	0,556
EMPLOYMENT CONTRACT	31620,41	0,549
CCAA	31636,444	0,545
BANK	31483,781	0,539
EMPLOYMENT SENIORITY	31683,59	0,536
SALARY INDEX	31751,06	0,511
AFFORDABILITY	31798,28	0,509

Before choosing the variables, it is necessary to look at the correlation between them. In the case of the variables education level and profession both are correlated, so the introduction of one of them in the model implies the redundancy of the other. We are left with the variable profession.

Once this is done, we choose these variables: breaches index, incidences in the system index, indebtedness, customer seniority, debit balance, asset balance, solvency, purpose, loan to value and profession.

We can see the ROC curves in the graphic below:



Now we do a logistic regression with all the chosen variables to obtain the scoring model.

The goal of logistic regression is to correctly predict the default index using the most parsimonious model. To accomplish this goal, a model is created that includes all predictor variables that are useful in predicting the response variable. Logistic regression can test the fit of the model after each coefficient is added or deleted by the stepwise regression.

It should be noted that variable solvency is not mathematically significant for the model but we include it because it makes business sense.

In this table we can see the scoring model variables and their estimated parameters beta:

VARIABLES	β
BREACHES INDEX	2,5409
INCIDENCES IN THE SYSTEM INDEX	0,8949
INDEBTEDNESS	-0,2865
CUSTOMER SENIORITY	-0,00179
DEBIT BALANCE	0,00002
ASSET BALANCE	-4,586E-7
SOLVENCY	-0,0439

VARIABLES	β
PURPOSE	1,2210
	-1,2622
	0,0412
LOAN TO VALUE	-0,170
	0,0432
	0,0536
	0,1006
	-0,0274
PROFESSION	0,1431
	0,3676
	-0,5107

The logistic regression model is formulated as follows:

$$P(\text{DEFAULT} \mid \text{AGE, SOLVENCY...}) = \frac{1}{1+e^{\beta_1 x_1 + \cdots + \beta_n x_n}} =$$

$$\frac{1}{1+e^{2.5409x_1+0.8949\,x_2-0.2865x_3-0.00179x_4+\cdots+0.3676x_{17}-0.5107x_{18}}}$$

With:

x_1 : BREACHES INDEX
x_2 : INCIDENCES IN THE SYSTEM INDEX
x_3 : INDEBTEDNESS
x_4 : CUSTOMER SENIORITY
x_5 : DEBIT BALANCE
x_6 : ASSET BALANCE
x_7 : SOLVENCY
x_8 : PURPOSE - Investments
x_9 : PURPOSE - Refinancing
x_{10} : PURPOSE - Consumer credit
x_{11} : LOAN TO VALUE - [0 0,4]
x_{12} : LOAN TO VALUE - [0,4 0,8]
x_{13} : LOAN TO VALUE - [0,8 0,9]
x_{14} : LOAN TO VALUE - [0,9 1]
x_{15} : LOAN TO VALUE - [1 1,5]
x_{16} : PROFESSION - Civil servants and graduates
x_{17} : PROFESSION - Freelancers.
x_{18} : PROFESSION - Others.

It can be seen from the data above that all the parameter signs make sense. For example, the beta of the variable breaches index is positive, which means that the probability of default increases when the variable grows. In the same way ,the beta of the variable solvency is negative which means that the probability of default decreases when the variable grows. As we expected.

3. ERROR QUANTIFICATION

3.1. Sources of model risk

Credit scoring models are based on historical data from a bank's portfolio (retail banking, individuals and SMEs) that serves to predict the likelihood of a customer defaulting on a new account. They provide a score of every binomial customer-operation to conclude whether they are "good" or "bad".

Within this framework, models can also be considered as a source of risk.
The possible adverse consequences (including financial loss) of decisions based on models that are incorrect or misused is called **model risk**.

While fitting the proposed scoring model we have identified three main eventual sources of error:

1. Data quality.
2. Model specifications.
3. Model usage.

The first source of risk refers to low quality data within the dataset as in the case of missing values, outliers, discontinuous time series, meaningless values (out of range values).

On the other hand an incorrect model specification can lead to worse model effectiveness rate, as the model has less predictive power in distinguishing between "good" and "bad" customers.This is the case when relevant variables haven't been taken into account (e.g. variables with a low p-value).

The third source of risk is related to the model usage. A classical example consists in a model fitted on a target variable and then used to predict a different one.

Prevent or mitigate model risk is possible. Data pre-processing is the right approach ensuring data quality. It allows the modeller to fill missing values, get rid of outliers values, fill incomplete data series or analyse data consistency.
Regarding model specifications we cannot eliminate the risk but we can try to assess it. It's possible to estimate how much the model effectiveness worsens as we introduce irrelevant variables and get rid of relevant ones.

Finally, in order to preserve the correct model usage, re-training represents an effective solution mitigating model risk.

3.2. Cutting point test

At this stage we focus on the model risk quantification. The dataset is divided in 10 subsets through a stratified sampling process. The partitioning variable is the default indicator variable. The final goal is avoid to train and test the model on the same dataset.

Learning the parameters of a prediction function and testing it on the same data is a methodological mistake: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data. This situation is called **overfitting**.

A solution to this problem is a procedure called cross-validation.
In the basic approach, called **k-fold cross-validation**, the training set is split into **k** smaller sets.

A model is trained using **k-1** of the folds as training data; the resulting model is validated on the remaining part of the data (it is used as a test set to compute a performance measure).This approach can be computationally expensive, but does not waste too much data.

At this stage we assess how our model performs if the **cutting point** by which it scores is modified.The cutting point is the barrier level that determines if the probability assigned by the model to a customer classifies it as a default or not.

If the model probability is lower than the cutting point the customer is classified as "bad" otherwise as "good" (1 default 0 no-default). The SAS default cutting point is 50%.
A **10-foldcross validation** is performed and we calculate for each fold the Error Type 1,Error Type 2, Accuracy and Power.

The goal is analyse the indicators evolution across folds and assess if the model can generalize to different datasets.

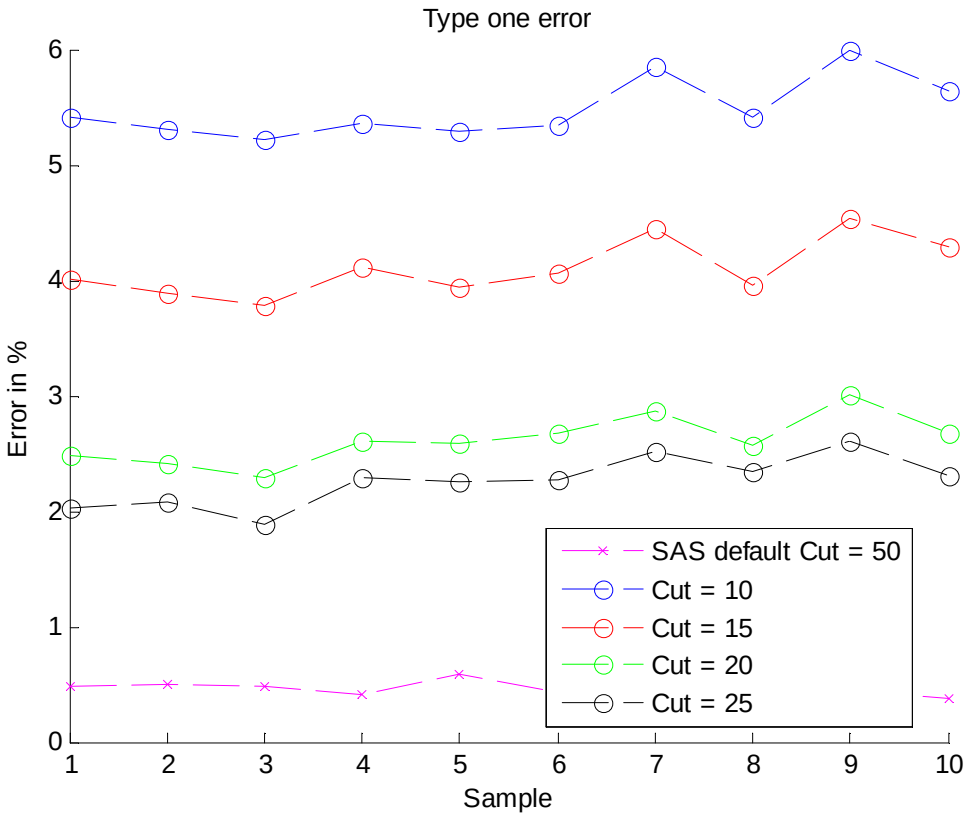
The confusion matrix (considering default as a positive event) is:

		ACTUAL STATUS	
		DEF.	NO DEF.
PREDICTED STATUS	DEF.	**TP**	**FP (Opportunity cost)**
	NO DEF.	**FN (Delinquency rate)**	**TN**

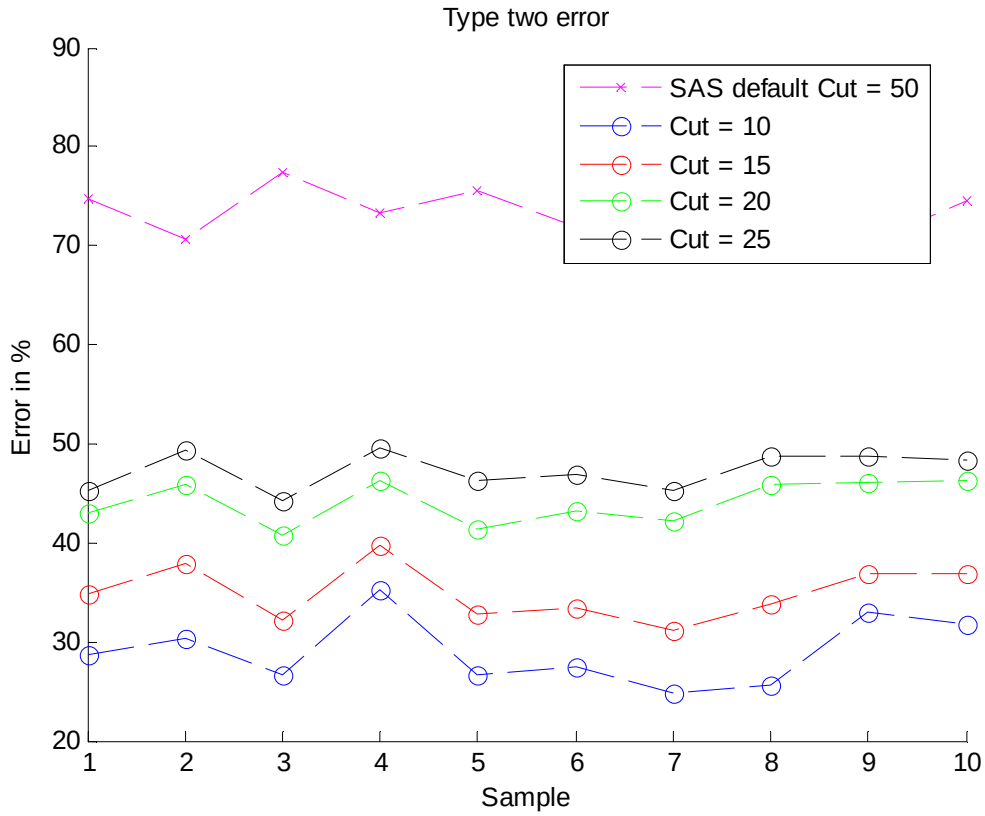
Type I error = FP / (FP + TN)
Type II error = FN / (TP + FN)
Power = 1 – Type II error
Accuracy= (TP + TN) / (TP+TN+FP +FN)

The base case is the 50% cutting point and the test considers cutting points of 10%, 15%, 20%, 25%:

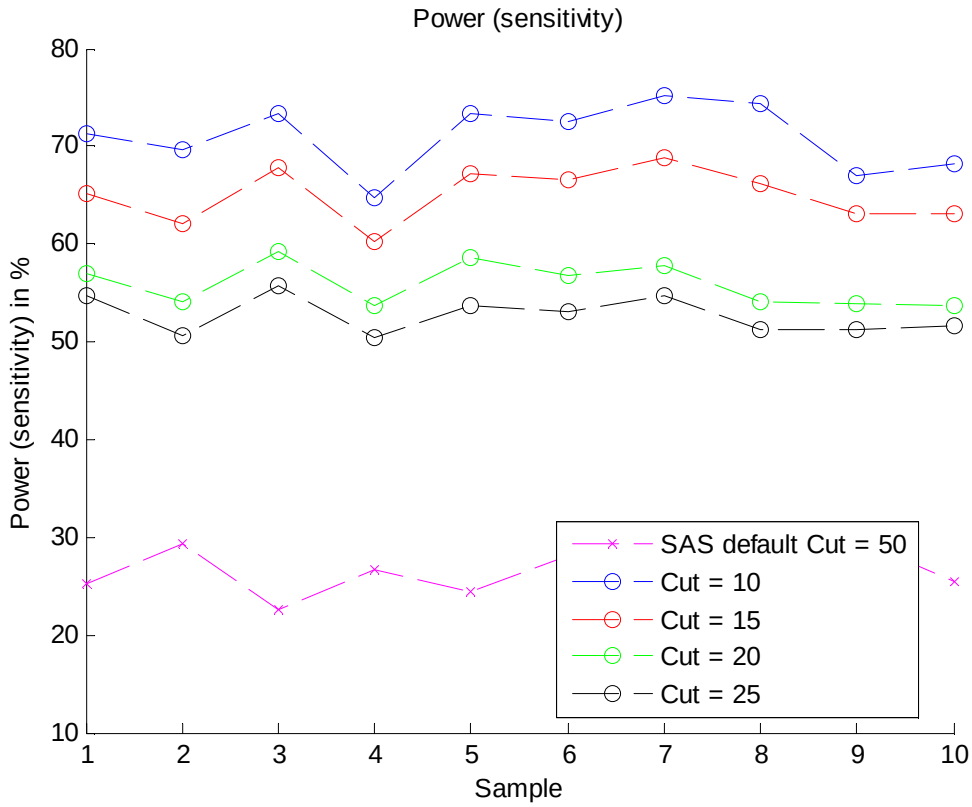
In business terms we refer to Delinquency rate as the number of customers classified as "good" that are defaults while Opportunity cost is the number of customers classified as default that are no default.



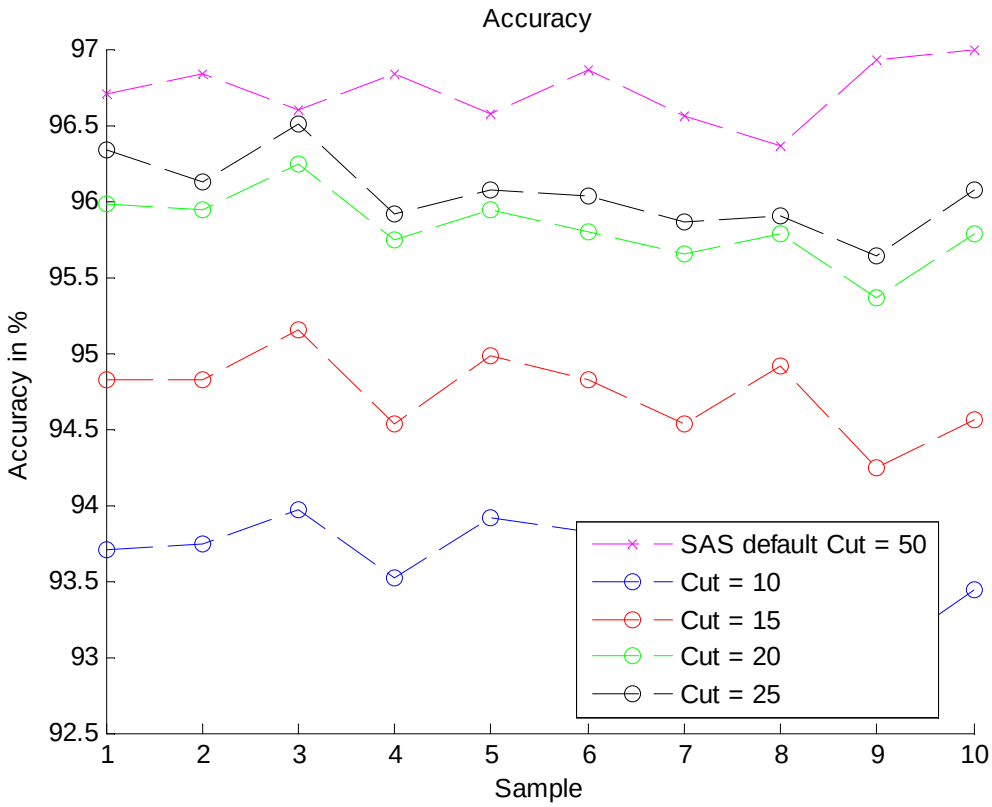
(1) Source: Own elaboration based on data supplied



(2) Source: Own elaboration based on data supplied



(3) Source: Own elaboration based on data supplied



(4) Source: Own elaboration based on data supplied

The evolution of the indicators across samples is almost smooth so we can conclude that our model can generalize to different datasets.

Regarding to the Type I and Type II errors in graphs **(1)** and **(2)** we observe that with a 50% cutting point the models exhibits a low Type I error but a high Type II error (very close to 75% on average).

If we consider a scoring parameter of 20% the Type I error rises only to 2.7% (on average) while the Type II error decreases to 45% (on average).

The decrease in Type II error is considerable if compared to the increase in Type I error so we conclude that a 20% cutting point is a good option in order to manage the trade-off between the Type I and the Type II errors.

Furthermore the model still has an acceptable power indicator of 55% and a high Accuracy rate of 95.5% on average.

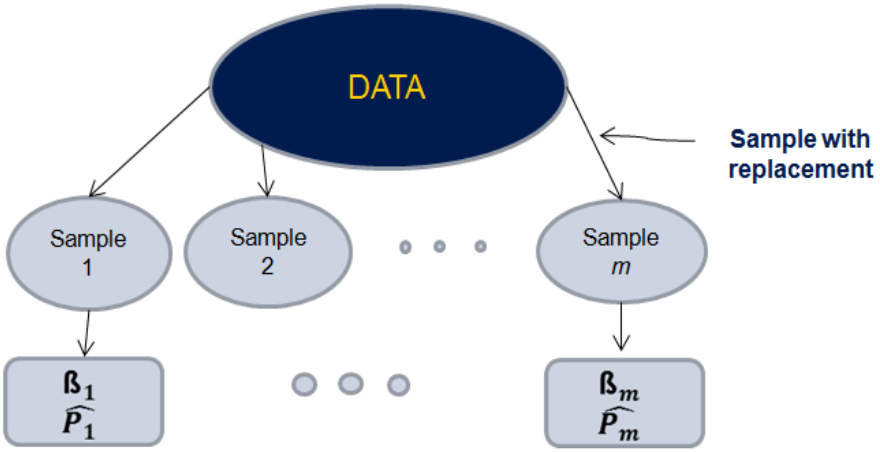
In accordance with this analysis the proposed scoring model has a cutting point of 20%.

3.3. BOOTSTRAP

Now we are going to talk about another method that helps us to quantify the model risk.

In general bootstrap provides a way to evaluate the empirical sampling distribution of parameter estimates (SEE, REFERENCE 2). This empirical sampling distribution can be used in similar manner to the theoretical sampling distribution.

How does the bootstrap work?



We are given an observed data set of size N. In our case we had a table with 100000 observations. We would like to point out that after the preprocessing of the data we are left with a few less observations let say n.

STEP 1. Resamplethe data with replacement, the size of the resample is equal to the size of the original data set n. This is called a bootstrap sample.

STEP 2. Beginning with the bootstrap sample, we run a logistic regression with the same variables as the best model that we have explained in a previous section. From the model we obtain parameter estimates beta and estimated probabilities for each observation using these betas.

STEP 3. Go to step 2 and repeat m times (we repeat this routine many times to get a more precise estimate of the Bootstrap distribution of the statistic).

The distribution of the m estimates of the betas represent the empirical sampling distribution. Making use of this empirical sampling we provide:

- Empirical confidence intervals for the betas. In order to achieve this, we take 5 and 95 percentiles of the empirical sampling distribution to form a 90% empirical confidence interval.

- The mean variation probability .

As a result we have:

	LOWER(5%)	UPPER(95%)
Intercept	3,28977317	3,736303477
PURPOSE_cat1	1,157832208	1,285968082
PURPOSE_cat2	-1,346228602	-1,1917576835
LOAN_TO_VALUE_cat1	-0,245616716	-0,086593108
LOAN_TO_VALUE_cat2	-0,044013649	0,11665272
LOAN_TO_VALUE_cat3	-0,085491791	0,201614398
LOAN_TO_VALUE_cat4	-0,022634659	0,226874892
PROFESSION1	0,031470726	0,291599965
PROFESSION2	0,214026569	0,502752614
PROFESSION3	-0,179159096	-0,02852693
PROFESSION4	-0,383910421	-0,249388838
INDEBTEDNESS	-0,362307138	-0,231471716
CUSTOMER SENIORITY	-0,002964746	-0,000570299
SOLVENCY	-0,307388807	0,295424897
ASSET BALANCE	-6,08E-07	-2,84E-07
DEBIT BALANCE	0,000012271	0,000021799
INCIDENCIDENCES	0,836160777	0,971189991
BREACHES INDEX	2,476731650	2,618133024

MEAN VARIATION PROBAB	0,33%

In the previous table we have an interval for each estimation of scoring model parameter, we could verify that the parameter estimates provided before lie within the limits as we expected. We do also have that the mean variation of the probabilities is 0.33%.

4. GLOBAL RISK MEASURE

We have also performed a global risk measure, this is a general formula that encloses all errors made in the model (quatifies the impact of model estimation on the model risk).

The global risk measure has been worked out combining some of the error measurements that the models provide. In particular we have made use of:

- Area under the ROC curve.
- Akaike information criterion (AIC).
- Type I error and type II error.
- Accuracy.

It is mandatory to standardise these measures (between 0 and 1), otherwise it would not represent any possible error.

We would like to point out that we have made the global risk measure to be 0 when we want to represent a model without error and 1 for a model which is totally wrong (making it is as simple as computing 1 minus the quantity that you have).

Once all the quantities have been standardised, we get the mean of all of them:

$$\frac{ROC + AIC + \cdots + ACC}{5}$$

In order to check that this proposed measure works, we have compared the best model that we got with another model where we introduce the variable civil status instead of purpose. The results are:

BEST MODEL:

ERROR MEASURE	VALUE	STANDARDISED VALUE
ROC	0.921	0.921
AIC	19068.761	0.7
TIE	2.6173	0.0261
TIIE	44.0952	0.4409
ACC	95.8252	0.9582

GLOBAL RISK MEASURE:

0.2376

PURPOSE ⟷ CIVIL STATUS:

ERROR MEASURE	VALUE	STANDARDISED VALUE
ROC	0.874	0.874
AIC	22678.55	0.7138
TIE	1.8933	0.0189
TIIE	63.8175	0.6381
ACC	95.7836	0.9578

GLOBAL RISK MEASURE:

0.3078

As we expected, the error increases when we change a variable in the best model that we had, so the measure that we have proposed seems to work well.

5. CONCLUSIONS

It is clear that banks and finantial institutions rely heavily on quantitative analysis and models in most aspects of financial decision making.

The expanding use of models in all aspects of banking reflects the extent to which models can improve business decisions. However, model risk may arise as a consequence of different sources of error. It is important to quantify this model risk since every entity should have sufficient resources to absorb the losses of its activity.

In order to manage this problem, we have developed a statistical methodology to address model risk. Particularly, we propose a global risk index to quantify the whole risk commited by a model.

References:

1. Carol Alexander, Mastering Risk, Volume 2: Applications, FT Press, 2001.

2. Bootstrapping article. Available at: http://www.columbia.edu/~mmw2177/LVcourse/bootstrap1.pdf (Accessed: 2 July 2014).