Practical Considerations for Testing the Cajamar Use Case

Sigve, Helge, Thomas, Ana, Ramon, Antonio

December 6, 2014

1 Cajamar: Test and evaluation

This section introduces the testing procedures for the Cajamar use cases. It builds on the general principles for evaluation already described in the previous sections of this report, and exemplifies how these principles can be employed in the setting of the credit evaluation application scenarios.

1.1 Use-case requirements

The test and evaluation procedures for Cajamar will be developed along the lines introduced in Deliverable 2.1 (D2.1): Instead of testing each use case separately, we utilize the notion of *application scenarios*. An application scenario is defined by a sequence of use cases that combined constitutes a full interaction procedure leading to a verifiable result. In D2.1 we defined two scenarios:

CAJ1: Prediction probability of default: The application scenario covers the first five use cases defined in Deliverable 1.2 (D1.2):

- UC1: Data reading and attribute construction
- UC2: Feature selection
- UC3: Model construction
- UC4: Model application
- UC5: Result checking and risk update

CAJ2: Low risk profile extraction: This application area uses the same model that is developed in the first scenario, but then progresses to use the model differently. It covers the following use cases:

- UC1: Data reading and attributes construction
- UC2: Feature selection
- UC3: Model construction
- UC6: Profile extraction

ID	Sub-phase	Description	Task(s)
CAJ.U1.O1	Interface	SQL queries should be efficient enough so that the whole process takes less than 3 hours.	8.2
CAJ.U2.O1	Interface	The feature selection should be efficient enough so that the whole process takes less than 3 hours.	4.3
CAJ.U3.D3	Testing	AUROC should be higher than 90%.	8.3
CAJ.U4.D1	Develop.	Model application should be efficient enough so that the whole process takes less than 3 hours.	2.3, 3.3, 4.1, 4.4
CAJ.U4.O1	Testing	Model should be able to evaluate daily about 5.6M clients.	2.3, 3.3, 4.1, 4.2
CAJ.U5.O1	Interface	The risk data update process should be efficient so that the whole process takes less than 3 hours.	8.2
CAJ.U6.O3	Testing	Expected benefits of a marketing campaign using obtained profiles should be 5% higher than with current methods.	8.3

Table 1: Testable requirements for the Cajamar use-case.

Requirements for the different use-cases were defined in D1.2. Most requirements are functional in nature, but some also introduce hard requirements that can be tested quantitatively. The latter are repeated in Table 1 for completeness. We note that the requirements center around three issues:

1. The whole process covered by Application scenario CAJ1, starting with SQL queries and ending with report generation must take less than 3 hours for the 5.6M clients (requirements CAJ.U1.O1, CAJ.U2.O1, CAJ.U4.D1, CAJ.U4.O1, and CAJ.U5.O1).

2. The prediction quality for Application scenario CAJ1, evaluated by AUROC, must be higher than 90% (CAJ.U3.D3).

3. The quality of the profiles generated in Application scenario CAJ2 is required to improve the benefit of at least 5% (CAJ.U6.O3).

1.2 Model and data characteristics

1.2.1 The data generation process

Both application scenarios use the same dataset, containing the defaulting behaviour of the Cajamar clients[1]. We now briefly describe the data generation process (the description is adapted from D2.1, where a more comprehensive description can be found). Please refer to Figure 1 for the timeline.

The dataset is created at time k, and contains a record for every client to be evaluated. Predictive variables refer here to the financial activity and payment behaviour of the customers in recent past as well as to their socio-demographic information, which usually does not change over time.

[1]The second application scenario will use only a subset of the total number of features.

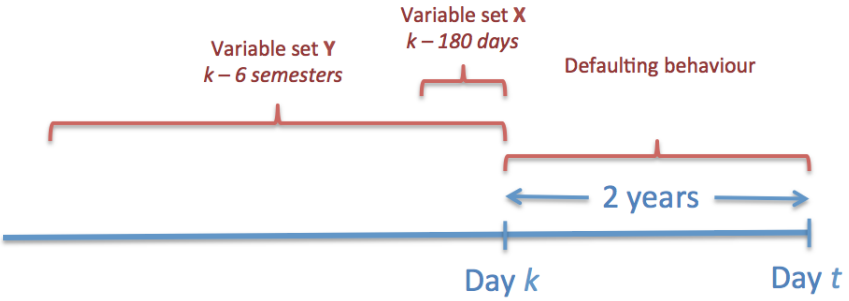

Figure 1: Time-line showing the generation of the data set. t refers to the present time and k corresponds to time $t - 2$ years. There are two disjoint groups of variables, denoted as $\mathbf{X}$ and $\mathbf{Y}$, with different past information considered, 180 days back (daily) and 6 semesters back (aggregated by semester), respectively.

Attributes denoted by $\mathbf{X}$ refer to the financial activity during the last 180 days. Examples of these features include "account balance" and "number of credit card operations". They usually change daily for a customer and are encoded by introducing a set of variables for each attribute – one for each day back from time k. Hence, the financial activity of a customer is specified by a number of variables equal to 180 times the number of attributes.

For others attributes, denoted by $\mathbf{Y}$, we are interested in information from the last 36 months. Examples of variables in this set include payments inside Cajamar (loans, mortgages, credits, etc.). The information from the last 36 months is grouped by semester, giving 6 summary variables per attribute that is considered. Finally, there are some static variables (mainly encoding socio-demographic aspects) denoted by $\mathbf{Z}$. These are not included in Figure 1 as they are not time-indexed.

The objective of the data analysis is to detect if a customer with profile $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ will default within the next two years. This corresponds to the class label in the dataset, *Defaulter*, and is determined by inspecting the user's behaviour from time k and 2 years into the future, i.e., in the period from k to t (see Figure 1). We obtain this information directly from Cajamar's databases simply by selecting the time k to be two years back in time, and thereby letting t be the current time. Note that, at the present time (time t), we have information of the *Defaulter* variable in the period of time from k to t. Thus, Defaulter$^{(k)}$ indicates if at some point in this period the customer was a defaulter.

The format of the data set for training/updating the model is depicted in Table 2. Each record contains the values for all predictive variables and a class variable. The class variable is labelled as *non-defaulter* only when there is no defaulting in the period from k to t (2 years).

1.2.2 The generated dataset

The existing data set was generated at t equal to December 31$^{\text{st}}$ 2013, thus simulating the calculations as if the AMIDST system was run two years before that (k corresponds

	Days			Semester				
Time t	$\mathbf{X}^{(k-180)}$	$\dots$	$\mathbf{X}^{(k-1)}$	$\mathbf{Y}^{(k-6)}$	$\dots$	$\mathbf{Y}^{(k-1)}$	$\mathbf{Z}$	Defaulter$^{(k)}$
Client$_1$								
$\vdots$								
Client$_n$								

Table 2: Three groups of attributes $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are distinguished according to the past information required. Current time is denoted as t. The data set is built at time t with $k = t - 2$ years.

to December 31st 2011). The dataset includes all customers who have been a client of Cajamar in the period k and up to day t, corresponding to $n = 4.5$M customers.

Every member of the dataset has been classified as either non defaulter or defaulter (no missing entries). Customers can have missing attribute values in their description. In particular, some of the clients were not clients in the whole three year period before day k. If a customer was not associated to Cajamar at some point in time $k - 6$ semesters, there will be missing values for some of the variables in $\mathbf{Y}^{(k-6)}$. Formally, every member i of the dataset has a vector of explanatory variables denoted by $\mathbf{W}_i = \{\mathbf{X}_i^{(k-180)}, \dots, \mathbf{X}_i^{(k-1)}, \mathbf{Y}_i^{(k-6)}, \dots, \mathbf{Y}_i^{(k-1)}, \mathbf{Z}_i\}$, potentially with some missing values (encoded using special codes). In total, each customer can be described using 7036 variables, that is, $180\,|\mathbf{X}| + 6\,|\mathbf{Y}| + |\mathbf{Z}| = 7036$, where $|\cdot|$ is the number of variables in each group.

Of all the customers in the dataset, 76% are considered to be of no risk because they do not have a loan in the bank, approximately 20% of the customers were exposed but did not default, and 3.79% of the customers defaulted in the two-year period of interest.

This data set will be used both for training and testing purposes. The test-set is defined by randomly selecting 20% of the customers. For reproducibility, a documented procedure including a fixed random seed is used to select the customers in the test set.

1.3 Predictive performance: test and evaluation

1.3.1 Application scenario 1

The goal of the first application scenario is to determine if a customer is going to be a defaulter after two years. This corresponds to a classification problem, where the class variable is denoted as Defaulter^k in Table 2.

The approach of the AMIDST project is to calculate $r_i = P(\text{Default}_i^k \mid \mathbf{w}_i)$ for each customer i. These quantities update the risk table in the system (see Table 3). If, at some point, the probability of default of a customer rises above a predefined threshold, the bank may take preventive actions to reduce the risk of defaulting by this customer.

According to requirement CAJ.U3.D3, the risk prediction quality should be assessed using the area under the receiver operating characteristic (ROC) curve. The classification rule is that a customer i is classified as a defaulter if $r_i \geq C$ for some

Time t	Risk of being defaulter
Client$_1$	r_1
$\vdots$	$\vdots$
Client$_n$	r_n

Table 3: Risk for the bank customers where r_i represents the probability of being defaulter for customer i.

constant C, and the ROC curve is computed by plotting the rate of true positives against the rate of true negatives for various choices of C. The requested area can then be found directly, and should, according to the requirement, be larger than 0.90.

There are two main issues with the outlined approach:

Changes in the economic climate: A Cajamar customer's chance of defaulting is to some extent determined by external factors, like the economic climate in Spain. During the economic crisis, the rate of defaulters was significantly higher than what was observed prior to the onset of the crisis. The data set we work with, corresponds to the period of the crisis, and it is natural to expect that the relationships found by the model are optimized for that economic climate. The AUC criteria is chosen to remedy global effects that affect all customers in a similar way, but we can still not guarantee that the model will perform at a similar level for a fundamentally different economic climate. In some cases these differences in the economic climate might be overcome by some of the socio-economic variables, for instance, a civil servant's personal economy with a permanent job should be less influenced by the economic climate than a casual/seasonal worker.

Stable versus volatile climate: Customers in the training and test sets are per definition from the same time period. Learning from the training data, we will therefore be able to detect the economic climate to which the customers will be exposed (e.g., simply by detecting the fraction of defaulters in the training data). If put in production, the predictions the AMIDST model is asked to make will be about the future, i.e., shifted two years in time when compared to the training data. This is not a problem if the economic climate is static or only slowly varying, but will be disastrous if, for instance, AMIDST is asked to make predictions about future customers immediately before the onset of a new economic crisis.

To partly account for these shortcomings of the test procedure we have collected another dataset with k equal to December 31$^{\text{st}}$ 2013, and where the correct class labels will be discovered two years later (December 31$^{\text{st}}$ 2015). As of today, the class labels for this dataset are unknown, but will be supplied at the end of 2015, and will therefore be available for the formal testing in Task 8.3. An AUC of more than 90% when using this new dataset for testing (and the original dataset for training) would be seen as a strong indication of the applicability of AMIDST in production.

1.3.2 Application scenario 2

Currently, a marketing campaign in Cajamar involves two steps:

- The first step is conducted by the marketing department, and results in a list of candidate customers. The list contains clients that have a high probability of signing what is offered to them (for instance a credit card).
- The second step, conducted by the risk modellers, is to filter out clients that are risky in terms of defaulting.

The task described in Use case 4 is to find relevant users profiles. The profiles can, for instance, be used to conduct marketing campaigns. The profiles should contain customers that are likely to be non-defaulters, and cover only attributes that are found to be relevant by the domain experts. It is required (CA.U6.O3) that the expected benefits of a marketing campaign using the obtained profiles should be 5% higher than with current methods.

Direct quantitative evaluation of the AMIDST-generated profiles is difficult to perform in a formal way mainly for two reasons. The first reason is that the application scenarios generate a *user profile* and not a *set of users*. We cannot value a profile in itself; it is the application of the profile to generate user sets that can potentially be monetized. The second reason is that the AMIDST profile defines users that are not likely to default, not users that are likely to sign a contract (and therefore not necessarily users who are valuable as marketing objects). For instance, it seems natural to expect the AMIDST profile to prefer solvent customers living in their own homes without any mortgage and with a sizeable cash-account. On the other hand, a customer like that may not be relevant to target for a campaign selling small-sized cash-loans without security requirements. This leads to a two-tier evaluation of the profiles, first considering the profiles as generators of profitable marketing campaigns, next as a way to evaluate the ability to find customers that will not default.

1) EVALUATION OF A CAMPAIGN'S PROFIT: Cajamar currently uses theoretical measurements for evaluating marketing campaigns. Let $s_i = P(\text{Signs}_i \mid \mathbf{w}_i)$ be the probability that a customer i signs on an offer presented to him (calculated by an existing system used by Cajamar) and, as before, let $r_i = P(\text{Default}_i^k \mid \mathbf{w}_i)$ be the probability that customer i defaults on a loan within the next two years (given that the offer is accepted). Furthermore, let γ_i be the net present value for the bank of that offer given that the customer does not default, and Γ_i the cost of the offer if a customer ends up in defaulting. c is a fixed indirect cost of the campaign. Then, the loss of a customer would be

$$L(\mathbf{w}_i, s_i, r_i) = \begin{cases} s_i r_i \Gamma_i + (1 - s_i)(1 - r_i)\gamma_i + c & \text{if a loan is offered;} \\ s_i(1 - r_i)\gamma_i & \text{otherwise.} \end{cases} \tag{1}$$

We therefore propose that the profile extraction is evaluated as follows:

1. A marketing campaign is selected, and the set of customers contacted are listed. The set of customers is called $\mathcal{C}$.
2. The AMIDST system is used to generate a profile for non-defaulting customers, and a fixed number of customers fitting the profile (comparable to the number of elements in $\mathcal{C}$) are selected. The marketing department

selects a subset of the customers in this set based on their probability to contract. Call this reduced set of customers $\mathcal{A}$. Note that the set $\mathcal{A}$ now defines a fictitious campaign (chosen by AMIDST) and has not been employed in a real campaign.

3. The two sets $\mathcal{C}$ and $\mathcal{A}$ are compared qualitatively and quantitatively (using the theoretical measure in Equation (1)).

In Equation (1) all customers that are not included in the campaign will contribute with the loss $s_i(1-r_i)\gamma_i$. In practice, contributions will only be collected from the set of customers that is selected by either method, i.e., the set of customers in $\mathcal{C} \cup \mathcal{A}$ because a customer selected by neither system (i.e., not in $\mathcal{C} \cup \mathcal{A}$) will contribute equally to the loss of each method, and is therefore not helpful to establish the difference between them.

2) EVALUATION OF DEFAULTING BEHAVIOUR: A drawback of this approach is that the loss-function rests upon (theoretical) probabilities $P(\text{Default}_i^k \mid \mathbf{w}_i)$ and $P(\text{Signs}_i \mid \mathbf{w}_i)$. Due to the introduction of $P(\text{Signs}_i \mid \mathbf{w}_i)$ in the loss function, we cannot in general guarantee that the system that is best at predicting the defaulting behavior will obtain the lowest loss. To target the quantitative requirement without using the theoretical probabilities we therefore also propose to utilize the AMIDST risk prediction capability from application Scenario 1 directly in the marketing setting, where the following procedure will be performed:

- Select a historical marketing campaign that is at least two years old, and remove all customers that did not sign. The remaining set of customers is called $\mathcal{C}$.
- Filter out clients that the AMIDST system deem too risky. The set of customers is called $\mathcal{A}$. Note that $\mathcal{A} \subseteq \mathcal{C}$.
- Calculate the empirical loss of the set $\mathcal{A}$ compared to that of $\mathcal{C}$. The requirement is that the loss of $\mathcal{A}$ should be at least 5% lower than the loss of the set $\mathcal{C}$. Note that for a direct comparison, only the difference in the two sets, $\mathcal{C} \setminus \mathcal{A}$, will contribute, and the costs of applying the AMIDST solution will be γ_i for the customers that do not default and $-\Gamma_i$ for those that do.

It should be noted that this procedure only evaluates the AMIDST system's ability to remove poor customers from the list of customers that were included in the original campaign, as we are unable to quantify the effect of AMIDST potentially wanting to send marketing material to customers that were not selected for the historical campaign without using the theoretical construct of Equation (1). The two evaluation approaches must therefore be seen in combination to give the full picture.

1.4 Run-time performance: test and evaluation

1.4.1 Application scenario 1

According to the requirement procedure, the full process starting with SQL statements and ending with validation of the new risks should be completed in no more than three hours (CAJ.U1.O1, CAJ.U2.O1, CAJ.U4.D1, CAJ.U5.O1).

The AMIDST solution will be installed in a server (IBM System x3690 X5) with the following characteristics:

- 2 Intel Xeon 10C Processors (Model E7-2870 130w 2.40GHz/30MB)
- 256GBRAM,16x16GB(1x16GB,4Rx4,1.35V)PC3L-8500CL7ECCDDR3 1066MHz LP RDIMM
- 2 internal disks SAS IBM 146 GB 2.5in SFF Slim-HS 15K 6Gbps SAS HDD RAID, one of them hot swap. 10 Disks IBM 600GB 2.5in SFF 10K 6Gbps HS SAS HDD.

This server is mainly used by credit risk models and marketing models departments. Data will be obtained via SQL queries. The current Information Center database management system is Oracle, but is being changed to a Teradata solution.

Learning the AMIDST model, however, requires an iterative process that is performed until convergence, and whose number of steps, and hence time, might vary due to random initialization. One way to enforce that the 3 hour upper bound is not violated, is to specify a *timeout* counter for the learning algorithm. Ideally, there should be enough time for the learning algorithm to reach convergence, and the timeout counter should be only included as a safe mechanism.

In order to provide a reliable estimation of the average time employed by the learning process and its expected performance, the following two graphs will be plotted:

- A first graph whose x-axis shows the number of tests and y-axis the total time employed by each of the experiments. Apart from the mean, a confidence interval at e.g. a 99% confidence level, will be provided to guarantee that the expected time will be included among this interval with a margin of error of 1%.
- A second graph in which the x-axis corresponds to the number of iterations/time and y-axis to the performance of the learning algorithm (indicating how close it is to convergence, e.g. lower bound in variational Bayes).

Joint interpretation of the two graphs will give us knowledge about the expected running time and performance of the process. We want to provide a mechanism to ensure that with a level of confidence of 99%, the learning algorithm will converge at the desired time limit. In the worst case, the algorithm will provide a valid outcome that might not be optimal. Note that the same analysis should also be performed for any other stochastic algorithms utilized in the process of the application scenario.

1.4.2 Application scenario 2

There are no specific run-time requirements for this application scenario. Specifically, it is stated that *"[. . .] execution time of this process is not relevant because the marketing campaigns are not launched so frequently."*.