Practical Considerations for Testing the Cajamar Use Case

Sigve, Helge, Thomas

November 21, 2014

1 Cajamar: Test and evaluation

This section introduces the testing procedures for the Cajamar use cases. Bla bla. Bla bla.

1.1 Use-case requirements

The test and evaluation procedures for Cajamar will be developed along the lines introduced in Deliverable 2.1: Instead of testing use cases separately, we utilize the notion of *application scenarios*. An application scenario is a 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 use scenarios:

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

- UC1: Data reading and attributes 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, and introduce functionality requests into the system, 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 takes 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 are evaluated through an improved benefit of at least 5% (CAJ.U6.O3).

1.2 Model and data characteristics

[[The following is from D2.1. Some slight adaptions, as I have removed all about the "Evaluation set". Should be made more to the point?]]

1.2.1 The data generation process

Both application scenarios use the same dataset, showing the defaulting behaviour of the Cajamar clients. 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.

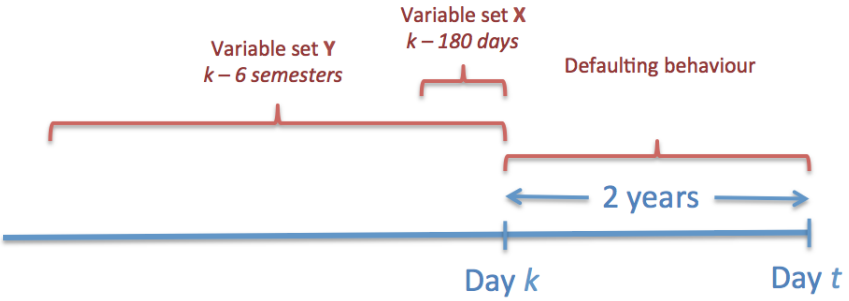

Figure 1: Time-line showing the generation of the data set. t refers to the present time and k corresponds to time $t-2$ years. There are two disjoint groups of variables, denoted as $\mathbf{X}$ and $\mathbf{Y}$, with different past information considered, 180 days back (daily) and 6 semesters back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

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.

The attributes denoted $\mathbf{X}$, for which financial activity during the last 180 days is considered. Examples of these features include "account balance" and "number of credit card operations". These attributes usually change daily for a customer, so they are encoded by introducing a set of variables for each attribute, one for each day back from the current time t. 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 $\mathbf{Y}$, we are interested in information from the last 36 months. Examples of variables here include payments inside Cajamar (loans, mortgages, credits, etc.). The information from the last 36 months are grouped by semester, giving 6 summary variables per attribute that is considered. Finally, there are some static variables (mainly includes socio-demographic variables) 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 $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ defaults within the next two years. This 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 back in time, and thereby letting t be the current time. We use the variable *Defaulter*. 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 full 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).

	Days			Semester				
Time t	$\mathbf{X}^{(k-180)}$	...	$\mathbf{X}^{(k-1)}$	$\mathbf{Y}^{(k-6)}$	...	$\mathbf{Y}^{(k-1)}$	$\mathbf{Z}$	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.

1.2.2 The generated dataset

The 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 coresponds to December 31*st* 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 where not clients in the whole three year period before day k. If a customer was not associated to Cajamar at time $k - 6$ semesters e will not have direct access to the variables $\mathbf{Y}^{(k-6)}$. However, Cajamar will manually fill in all relevant missing values using other data sources. Formally, every member i of the dataset therefore has a vector of explanatory variables denoted by $\mathbf{W}_i = \{\mathbf{X}_i^{(k-180)}, \ldots, \mathbf{X}_i^{(k-1)}, \mathbf{Y}_i^{(k-6)}, \ldots, \mathbf{Y}_i^{(k-1)}, \mathbf{Z}_i\}$ without missing values. In total, each customer is described using 7036 variables.

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 wee exposed but did not default, and 3.79% of the customers defaulted in the two-year period of interest.

The data set will be used both for training and testing. The test-set is defined by randomly selecting 20% of the customers. For reproducibility, a fixed random seed of

[[WHAT IS IT]]

was used to select the customers in the test set.

1.2.3 The models

[[SHOULD WE TAKE SOMETHING FROM D2.1 HERE, OR SKIP IT? I AM NOT SURE IF THE MODELS ARE OF RELEVANCE.]]

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 customers is going to be a defaulters after two years. This corresponds to a classification problem, where the

class variable is denoted Defaulter^k in Table 2.

The approach of the AMIDST project is calculate $r_i = P(\text{Default}_i^k \mid \mathbf{w}_i)$ for each customer i. These quantities are afterwards 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.

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

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

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 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 is according to the requirement to be larger than 0.90.

There are three main issues with the outlined approach:

Relevance of the data: All the customers' data made available to the project for training and testing belong to people already associated with Cajamar, and their economic status (e.g., regarding whether or not they have obtained a loan) is a consequence of Cajamar's current system being employed to remove the customers not deemed appropriate. The dataset we have access to may therefore not reflect the distribution of customers that approach Cajamar in the first place (and which AMIDST will be exposed to if employed in production).

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 crises the rate of defaulters was significantly higher than what was observed prior to the onset of the crisis. The data set we work with is from 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 fundamentally different economic climate.

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 AMIDST are 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 outset 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 31st 2013, and where the correct class labels will be discovered two years later (December 31st 2015). 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.

[[Remove this part, or will we have the extra data? I am writing that the data has already been generated just to comply with the "Data available from Day 1"-requirement from our friends in EU.]]

1.3.2 Application scenario 2

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 (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 to conduct marketing campaigns. The profiles should contain customers that are like 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 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. The first issue is that the application scenarios generates *user profiles* and not *set of user*. We cannot value a profile in itself; it is the application of the profile to generate user sets that potentially can be monetized. The second problem is that the AMIDST profile defines users that are not likely to default, not users that are likely to 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 sizable 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. We therefore propose that the project extraction is evaluated qualitatively as follows:

- A historical marketing campaign is selected, and the set of customers contacted are listed. The set of customers is called $\mathcal{C}$.
- 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 $\mathcal{A}$.
- The two sets $\mathcal{C}$ and $\mathcal{A}$ are compared qualitatively.

To target the quantitative requirement, we also propose to utilize the AMIDST risk prediction capability from application scenario 1 in the marketing setting, where the following procedure will be performed:

- Select a historical marketing campaign that is at least two years old. The set of customers is called $\mathcal{C}$.
- Obtain the empirical risk of the campaign during the two year period (see Section XX).

[[Sigve will write about empirical loss etc in another part of the overall document.]]

- Filter clients that the AMIDST system deems too risky. The set of customers is called $\mathcal{A}$. Note that $\mathcal{A} \subseteq \mathcal{C}$.
- Calculate the empirical risk of the set $\mathcal{A}$. The risk of $\mathcal{A}$ should be at least 5% lower than the risk of the set $\mathcal{C}$.

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 campaign, and we are not able to determine the effect of AMIDST potentially wanting to send marketing material to customers that were not selected for the historical campaign.

[[From Ramon:

- **Another thing that can be done is to estimate the benefit of clients that AMIDST would have included (using expected income and expected loss) and add it to the benefit of the virtual AMIDST campaign.**

Should we add that?]]

[[Rest of the section is Sigve's text. I propose that we remove all the math into the start-up-sections and just use established stuff like "Empirical risk" here. Not sure if all of Sigve's points have been included in the "for dummies" - version I have written. Must be checked.]]

There are more opportunities with testing the second step. After performing step one the test data set is reduced to a set of clients with a high contracting probability (i.e. above a certain level). Let the clients in this data set be (x_i, y_i), where y_i is either default or not default and x_i is a vector of explanatory variables. It makes sense to compute AUC for both the current method and the Amidst method to compare the two methods. This comparison will say something about the ability of the filter to take out defaulters, while keeping the non defaulters. However, we must assume that the real probability of contracting $P(\boldsymbol{x_i})$ is completely independent of whether the client will actually default or not.

Moreover, in Delivery 1.2, it is required that the benefit of a AMIDST induced marketing campaign should be more than 5 percent higher than a normal campaign. In

order to discuss such a requirement we have to iintroduce a function that describes the financial loss of a certain classification rule compared to a classification rule that make no mistakes.

In this paper, we define the *loss function* as a real and lower-bounded function $L(x_i, h(x_i), y_i)$. It takes into account the explanatory variables for each client x_i, the predicted class $h(x_i)$ and the true class label y_i.

In the current system in Cajamar the classification rule is denoted h_{Current,L_1} and is defined by

$$h_{\text{Current},L_1}(x_i) = P_{\text{Current}}(\text{Default}_i \mid \boldsymbol{x_i}) \leq L_1 \tag{1}$$

where the probability for defaulting client i are $P_{\text{Current}}(\text{Default}_i \mid \boldsymbol{x_i})$. Here, L_1 is a chosen classification limit.

We let the cost of excluding client i that does not default as $c_i(0|1)$ and also the cost of including client i that does default as $c_i(1|0)$. Both costs are related to the size of the potential offer. Also, we make the assumption that the real probability of contracting $P(\boldsymbol{x_i}) = p$ is completely independent of whether the client will actually default or not. The loss function below is of interest

$$L(x_i, h_{\text{Current},L_1}(x_i), y_i) = \begin{cases} 0 & \text{for} \quad h_{\text{Current},L_1}(x_i) = 0 \quad \& \quad y_i = 0 \\ pc_i(1|0) & \text{for} \quad h_{\text{Current},L_1}(x_i) = 1 \quad \& \quad y_i = 0 \\ pc_i(0|1) & \text{for} \quad h_{\text{Current},L_1}(x_i) = 0 \quad \& \quad y_i = 1 \\ 0 & \text{for} \quad h_{\text{Current},L_1}(x_i) = 1 \quad \& \quad y_i = 1. \end{cases} \tag{2}$$

Notice that L is an array of $n \times 2 \times 2$ elements. Cajamar can estimate $c_i(0|1)$ and $c_i(1|0)$ for all clients in the database.

The empirical risk is found by averaging the loss function on the training set given by

$$R_{emp}(h_{\text{Current},L_1}, \boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} L(x_i, h_{\text{Current},L_1}(x_i), y_i). \tag{3}$$

It is now possible to calculate the empirical risk involved with using both the current filter and also the Amidst filter. It is therefore possible to estimate the ratio between the costs and therefore see whether there is more than 5 percent gain in using the Amidst default filter compared to using the current default filter. Notice that in terms of estimating this gain percentage it is not needed to estimate p. However, it could be estimated from the number that accepted the offer on an old campaign.

Calculating empirical risk on an old campaign

It is also possible to use an old campaign to test the improvement of using the Amidst default filter in addition to the current filter.

Consider an old campaign that was done more than two years ago. Even though costs and default/non defaults are known for all clients, the loss function is only known

on the clients that was targeted in that campaign. This makes this discussion complicated.

We will now consider the AMIDST induced marketing campaign as a binary classification problem with class variable y_i, which can take the values $\{0, 1\}$. Class one refers to non-defaulters that actually signs the contract and class zero refers to the rest of the clients. In a perfect campaign only the non-defaulters that actually signs the offer (for instance a credit card or a loan) are selected.

In the current system in Cajamar the classification rule is denoted $h_{\text{Current},L_1,L_2}$ and is defined by

$$h_{\text{Current},L_1,L_2}(x_i) = P_{\text{Current}}(\text{Default}_i \mid \boldsymbol{x_i}) \leq L_1 \,\&\, P_{\text{Current}}(\text{Contract}_i \mid \boldsymbol{x_i}) \geq L_2, \tag{4}$$

where the probability for defaulting and contracting for client i are $P_{\text{Current}}(\text{Default}_i \mid \boldsymbol{x_i})$ and $P_{\text{Current}}(\text{Contract}_i \mid \boldsymbol{x_i})$. Here, L_1 and L_2 are chosen classification limits.

We let the cost of excluding client i that actually would contract and not default as $c_i(0|1)$. This cost is related to the size of the potential offer.

Moreover, we let $c_i(1|0)$ be the cost of offering to client i, provided that he either would not take the offer or would default if he took the offer. Clearly, if client i was offered and contracted but defaulted, $c_i(1|0)$ is related to the size of the contract. Otherwise, $c_i(1|0)$ is only related to the cost of making the offer. The loss function below is of interest

$$L(x_i, h_{\text{Current},L_1,L_2}(x_i), y_i) = \begin{cases} 0 & \text{for} \quad h_{\text{Current},L_1,L_2}(x_i) = 0 \quad \& \quad y_i = 0 \\ c_i(1|0) & \text{for} \quad h_{\text{Current},L_1,L_2}(x_i) = 1 \quad \& \quad y_i = 0 \\ c_i(0|1) & \text{for} \quad h_{\text{Current},L_1,L_2}(x_i) = 0 \quad \& \quad y_i = 1 \\ 0 & \text{for} \quad h_{\text{Current},L_1,L_2}(x_i) = 1 \quad \& \quad y_i = 1. \end{cases} \tag{5}$$

Cajamar can estimate $c_i(0|1)$ and $c_i(1|0)$ for all clients in the database.

The empirical risk for the old campaign is not taking into account the financial loss related to excluding a number of clients that actually would have contracted and not defaulted. Said with other words, the empirical risk is not taking into account losses related to when $h_{\text{Current},L_1,L_2}(x_i) = 0$, and $y_1 = 1$. In such a calculation none of the $c_i(0|1)$s are used. The empirical risk will therefore be less than the true risk (which would take the above point into account).

A simple test is to use the Amidst toolbox to provide an additional filter related to default prediction on top of the old classification rule. Mathematically this is

$$\begin{aligned} h_{\text{Amidst filter},L_1,L_2,L_3}(x_i) = & P_{\text{Current}}(\text{Default}_i \mid \boldsymbol{x_i}) \leq L_1 \,\&\, P_{\text{Current}}(\text{Contract}_i \mid \boldsymbol{x_i}) \geq L_2 \,\& \\ & P_{\text{Amidst filter}}(\text{Default}_i \mid \boldsymbol{x_i}) \leq L_3. \end{aligned} \tag{6}$$

This calculation of empirical risk is biased by the same amount as the old method. It makes therefore sense to compare $R_{emp}(h_{\text{Amidst filter},L_1,L_2,L_3}, \boldsymbol{x})$ with $R_{emp}(h_{\text{Current},L_1,L_2}, \boldsymbol{x})$. The benefit of using the Amidst model as additional filter can therefore be quantified.

Questions:

1. Do you see any flaw in reasoning?
2. What more should we do?
3. What do you think about the profiling ideas?

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 complicated in no more than three hours (CAJ.U1.O1, CAJ.U2.O1, CAJ.U4.D1, CAJ.U5.O1).

[[Testing procedure/how strictly we will look at this must be evaluated.]]

[[Mention the hardware specified in R1.2, p. 9?]]

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."*.