Practical Considerations for Testing the Cajamar Use Case

There are two application scenarios here. The first one is prediction of whether a client will default within two years or not and the second is related to the benefit of a marketing campaign. For clarity we have added the description about the data set from delivery 2.1 without any modification.

1 Description from 2.1

1.1 Predicting probability of default

Our objective is to tackle the current limitations of the risk prediction problem by *daily* learning the predictive model and also updating the risk of default for every bank customer. Dependences among the variables will now be considered, as well as including all the variables in the analysis. With these changes, Cajamar plans to improve the quality of the prediction model by increasing the area under ROC curve significantly.

Therefore, the process will consist in building a *training set* as well as a set of customers to be evaluated, called *evaluation set* (see Deliverable 1.2 [?]). How these data sets are generated gives us some insights into the nature of this risk prediction problem (see Figure 1 for a better understanding):

- **Model evaluation data set:** This data is created at time t and contains a record for every client to be evaluated. Note that information about the predicted defaulting behaviour is missing at time t and it will be obtained after performing inference on the model. 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.

  There are attributes, denoted as $\mathbf{X}$, for which information during the last 180 days is considered. 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 as $\mathbf{Y}$, we are interested in information from the last 36 months grouped by semester. Therefore, similar to previous group of variables, 6 variables for each of these attributes will be considered. Finally, there are some other static variables, denoted as $\mathbf{Z}$, not included in Figure 1 as
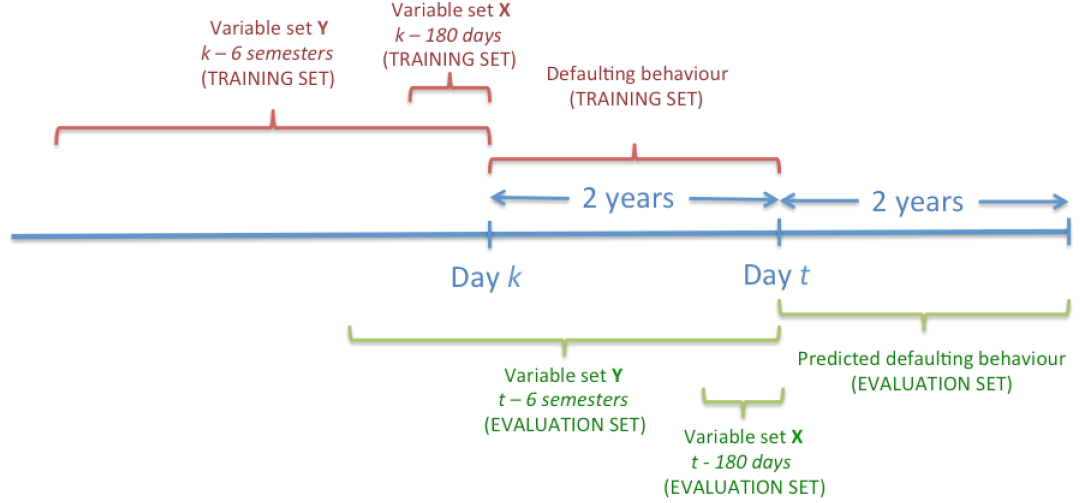
Figure 1: Time-line showing the generation of the evaluation (in green) and training (in red) data sets. t refers to the present time and k corresponds to time $t - 2$ years. Both in the training and test data sets, there are two disjoint groups of variables, denoted as $\mathbf{X}$ and $\mathbf{Y}$, with different past information considered, 180 days back (daily) and 6 semesters back (by semester), respectively.

they are not indexed over time. The data set for the evaluation of customers is depicted in Table 1.

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

Table 1: Evaluation data set at time t for all the clients. 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.

Thus, the objective is to compute the probability of defaulting within the following two years of each record from the evaluation data set, and afterwards update the risk table in the system (see Table 2).

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.

- **Model training data set:** This data set is also built at time t in a similar way as the evaluation data. It contains the same set of features as well as the target

2

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

Table 2: Risk table for the bank customers where r_i represents the probability of being defaulter for customer i.

variable *Defaulter* but with information referred to time k instead (two years back). Note that, at 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 she/he was a defaulter.

The data set for training/updating the model is depicted in Table 3.

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

Table 3: Training data set built at time t with $k = t - 2$ years. The notation for predictive variables is the same as in Table 1.

Table 3 shows the training data set where each record contains the values for all predictive variables and a class value labelled as *non-defaulter* only when there is no evidence of defaulting in the period from k to t (2 years).

2 Test and evaluation regime

The training set above is actually the data set that we are going to train on and also test on. This may be a source to confusion, but from now on the training set above will be defined as the dataset and the evaluation set is not even considered in relation to testing.

The only criterion for being a member of the dataset is that the member has been a client continuously from day k and up to day t. Every member of the dataset has a class label that is either non defaulter or defaulter. There are no missing values related to class labels. However, there are missing values related to attributes. In particular, some of the clients where not clients in the whole three year period before day k. Cajamar will manually fill in all relevant missing values. Formally, every member i of the dataset has a vector of explanatory variables that is denoted by $\boldsymbol{x_i} = \{\mathbf{X}_i^{(k-180)}, \mathbf{X}_i^{(k-1)}, ... \mathbf{Y}_i^{(k-6)}, \mathbf{Y}_i^{(k-1)}, \mathbf{Z}_i\}$.

We will divide this data set into a training set and a test set by a completely random process.

2.1 Requirements for the first use case scenario: default prediction

In the first use case scenario, it is required that the AUROC should be above 0.90.

The Bayesian network basically computes the probability $P(\text{Default}_i \mid \boldsymbol{x_i})$ and the classification rule is $P(\text{Default}_i \mid \boldsymbol{x_i}) \geq C$. The ROC curve is computed by basically plotting the rate of true positives against the rate of true negatives for various choices of C. AUROC is the area under the graph. (There is also another way of computing AUROC which do not involve computing the rate of true positives and rate true negatives for various C, but this is omitted in this discussion.)

Questions:

1. How many defaulters and non defaulters do we have on both the training set and the test set?

2. Can you go though all the information and make sure that it is correct.

2.2 Requirements for the second use case scenario: AMIDST induced marketing campaign

A marketing campaign in Cajamar involves two steps. The first step is to find clients that have a high probability of signing what is offered (for instance a credit card). The second step is to filter out clients that are risky in terms of defaulting. The testing regime involves testing each step separately.

The first step is difficult to test. We suggest that the Amidst profiler finds a list of a certain number of clients that are manually inspected by the marketing department in Cajamar. This list can for instance be compared to a list of clients with high contracting probability from a former campaign.

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

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

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?