The challenges of AI-automated banking decisions: Methodological considerations and large-scale implementation for overdrafts on current accounts in a bank

Pascal Damel – Hoai Le Thi An 1 – Van Tuan Pham

Laboratory L.C.O.M.S. – University of Lorraine

1 Professor, Institut Universitaire de France I.U.F.

THE RIA&MIT CHAIR

In training

presents its 3rd edition of the Workshop

- Artificial Intelligence and Organizational Impacts -

Abstract

The implementation of AI in banks is a technical and strategic challenge. The strategy is part of the productivity gains necessary to maintain profitability levels. The implementation of AI poses technical and methodological challenges. This article presents a large-scale experiment involving the implementation of AI in a bank to automate, using machine learning, all decisions regarding overdrafts on the current accounts of corporate and individual customers. In the initial phase, we were confronted with the complexity of banking information systems. We then present the main methodological choices available. In this article, we justify the methodological trade-offs made by following an original hybrid approach that combines the C.A.R.T. and S.V.M. methods. We also successfully proposed methodological advances in machine learning by creating a more suitable method, "C.S.B.-D.C.A.," for asymmetric data between overdraft approvals and refusals. The databases used in this study actually contain more approvals than refusals.

1. Introduction

Banks are increasingly using AI to automate certain banking decisions in order to reduce operating costs. Overdrafts are both a source of profit through commission fees and a source of risk in the absence of collateral on the amounts lent. The large volume of processing

The significant volume of manual processing of account overdrafts generates substantial operating costs. Each debit overdraft on a current account must be validated by the account manager. Many authorizations are straightforward and recurring, while others require in-depth and complex consideration based on multiple factors (customer relationship, outstanding customer debt, type of customer (legal entity or natural person), profitability/risk ratio, etc.).

Banks are embarking on this technological and methodological revolution with the following considerations.

What methodologies should be used? C.A.R.T. – machine learning – deep learning?

What technologies?

Should all decisions be automated, and under what conditions?

What are the consequences for the financial supply chain for businesses and the many organizational impacts for the company?

In this article, we will present the various considerations and advances that we formalized when implementing AI in overdraft decisions at a European bank for all corporate and individual customers, with nearly 450,000 decisions.

2. Epistemological aspects and database

This type of study fits perfectly within a positivist approach. Data are logical induction tools for understanding the decisions made.

1.1. The organizational challenges of bank overdrafts

Bank overdrafts on current accounts are a hot topic for consumers, banks, and European regulatory authorities. Current accounts are banking products on which banks earn a large part of their commercial margins. Commercial margins calculated using market rate methods [5] [6] on current accounts are significant because of their near-zero credit rates and the high fees charged when the authorized overdraft is exceeded.

The bank also has discretionary power when the customer exceeds the maximum overdraft limit. It can refuse the overdraft and consequently declare the customer insolvent. This decision is obviously a human one, with hierarchical levels depending on the type of customer and the amounts involved.

The challenge of automating overdrafts is a difficult and strategic decision. Poor automation can lead to legal and operational risks for the bank and its customers. Successful automation allows the bank to save many full-time equivalents.

In this respect, this implementation also concerns human resources. According to the bank's strategic discourse, the full-time equivalents gained by AI can be reallocated to additional commercial tasks that normally generate net banking income (NBI).

Bank overdrafts are complex to understand because customers are different. The bank must also take its decisions seriously, as current accounts do not generally offer any guarantee for the bank in the event of non-payment. These non-payments are reflected in the accounts as provisions for impairment.

Given the risks involved, banks generally limit automation decisions to modest amounts (a few hundred euros or even a few thousand, depending on the type of customer).

1.2. Databases

To begin this study, the first issue is to understand the complexity of the bank's information system and the regulatory constraints specific to the banking environment. As a regulatory constraint, we

complied with IFRS I.F.R.S. [1] 8 on customer segmentation. This segmentation is obviously legal, but it is also a "business" classification of customers. We also used the Basel regulations (Basel IV) which regulate the level of banks' capital based on the risks (credit, market, liquidity, operational) taken by the bank. These regulations offer risk calculation systems that have already been validated. Management control is also a source of information on customer profitability/risk. Of course, CRM also provides data. Obviously, the IT mechanisms that track overdrafts still need to be understood.

These databases are complex to obtain because the environment in which this data is stored is not all structured within a centralized information system. Banks have to deal with different IT systems, some of which date back to the 1970s (IBM400). Data is also compartmentalized into different IT chains to meet specific regulatory and internal requirements (accounting function, controlling function, back office functions, marketing, etc.). After a thorough audit of the quality of the available information, we can characterize customer information as follows:

– KYC (Know Your Customer) information on economic, legal, and tax criteria.

– Information on financial risks (Basel default probability – accounting information).

– Information on management control or management oversight (customer profitability history and marketing segment).

– Information on the operation of current and joint accounts (savings – number of loans).

– Technical information related to the decision to accept or refuse overdrafts on current accounts.

1.3. Which variables?

The thousands of available variables and impressive amounts of data do not allow decisions to be automated using global techniques such as deep learning.

To facilitate the work, the data was preselected. We selected expert variables based on consultation with managers from various analytical sources. We had 432,966 overrun decisions at our disposal, including 372,811 favorable decisions and 60,155 unfavorable decisions, distributed across the various customer segments. These variables were subjected to conventional statistical tests such as descriptive statistics and hypothesis testing (mean, variance, statistical distribution).

The purpose of these tests was to gain a better understanding of the decisions made during the credit approval process. We performed three statistical analyses: F-test (ANOVA), T-tests (means), and chi-square (seasonality test) to separate the variables between decisions. While the T-test explains the significance of the differences between group means, the main idea behind ANOVA for feature selection is to test the statistical significance of each feature's contribution to the response feature. The objective of these tests is to identify the relevant variables to be taken into account in the decision-making process and to exclude those that are not. These tests make it possible to reduce the variables from several thousand to around a hundred. The results of the analysis of the main data are presented in Table 1 for a specific customer segment in "premium retail."

[1] I.F.R.S. International Financial report standard

Table 1: Statistical tests

MAY2016 PR 2	MONTANT_AUTORISE_FORCAGE_modif	PROBABILITE_DEFAUT	GLOBAL_ASSETS	MT_REVENU_TOTAL	Nb_Doscred	Ancienc pt	V11_UNI	V32_UNI	V51_UNI
TESTS STATS									
Average (moment 1)	303.9911213	470.7232222	10248.90734	102.4294222	2.549951503	5469.103275	24.7311412	2600.959213	59.91093903
STD (moment 2)	271.5266572	1231.564632	21343.04395	184.215403	1.564434571	3210.374885	41.41672473	8037.941906	42.66304453
SKEWNESS (moment 3)	0.879321625	6.011177878	7.494997096	3.849055523	1.696794808	0.315022621	4.318051394	15.53728789	0.31929324
KURTOSIS (moment 4)	-0.345336407	42.0473417	84.093711	28.46291164	4.479026691	-1.065396421	20.91890929	389.4303894	-0.511660733
MIN	0.11	4	0	-43.338	1	3	1	-12465.9	0
MAX	290.89	9999	320680.3	2759.284	9	10743	232	229560.7	157.86
QUARTILE 1	80.2	34	877.76	2.86	1	2947	6	77.74728	20.906525
QUARTILE 2	215.91	65	3691.15	19.67	2	4978	14	879.9441	63.754
QUARTILE 3	471.54	572	11995	123.21	3	7971	25.5	3035.413	91.66355
COUNT	3389	3389	3389	3389	3093	3389	1551	3389	1564

MAY2016 PR 3	MONTANT_AUTORISE_FORCAGE_modif	PROBABILITE_DEFAUT	GLOBAL_ASSETS	MT_REVENU_TOTAL	Nb_Doscred	Ancienc pt	V11_UNI	V32_UNI	V51_UNI
TESTS STATS									
Average (moment 1)	283.6687514	788.2932417	6073.68921	84.91384192	1.949061662	4819.751432	37.28621908	1390.910367	75.2829515
STD (moment 2)	266.160674	1496.384694	18419.47266	187.6576311	1.085637995	2975.82914	52.76928633	4782.737372	37.72442434
SKEWNESS (moment 3)	0.965972659	4.485497987	8.951186537	8.712184571	1.287761022	0.628656423	4.295684582	7.309306364	0.106496565
KURTOSIS (moment 4)	-0.151551419	23.94825384	110.1251629	140.7635788	1.721764709	-0.614747825	27.97899361	80.35724207	0.989250543
MIN	0.26	4	0	-32.87	1	116	1	-12141.2	0
MAX	992.93	9999	252735.9	3580.18	6	10743	523	60257.51	200
QUARTILE 1	66.34	34	25	2.65	1	2307	12	-155.239	54.0931
QUARTILE 2	193.67	212	597.35	18.22	2	4331	21	235.2302	84.1313
QUARTILE 3	451.86	1145	3513.37	75.54	3	6464	38	1071.573	97.1146
COUNT	873	873	873	873	746	873	566	873	339

	MONTANT_AUTORISE_FORCAGE_modif	PROBABILITE_DEFAUT	GLOBAL_ASSETS	MT_REVENU_TOTAL	Nb_Doscred	Ancienc pt	V11_UNI	V32_UNI	V51_UNI
F TEST DISPERSION	0.465535613	4.88274E-14	1.06797E-07	0.482385422	1.21964E-31	0.005570852	5.12978E-13	2.19339E-67	0.005037877
T TEST ESPERANCE	0.045332551	9.20114E-09	9.40428E-09	0.013695777	1.81011E-33	1.87564E-08	3.90387E-07	1.45433E-08	7.6986E-11

The main variables selected for this customer segment at the 10% threshold relate to account history, credit risk probabilities, the number of outstanding loans and bank commitments, the length of the business relationship, assets under management, customer income, etc.

Bank overdrafts are more acceptable:

• when the probability of bankruptcy is low (probabilities are calculated using logistic regression methods).

• when the customer has assets (other accounts and investments).

• when the customer is active (an active customer is a profitable customer in terms of management control).

• when they are strongly committed to the bank. It is more difficult to refuse an overdraft when the company has to repay its debts to the bank.

• when the customer has been with the bank for a long time. It is easier to accept an overdraft for a long-standing customer.

• When the account's operating indicators are good (V11_UNI – V32_UNI: average account balance – number of days of excess, etc.).

Statistically, we did not find a significant seasonal effect at the 5% threshold with the Chi-square test on overdraft decisions, regardless of the type of customer. This aspect will be developed in another article exploring the impact of AI on the financial supply chain of companies.

2. Possible methods and justifications

We have previously ruled out deep learning for technical reasons and due to the dimensions of the variables.

In these positivist approaches, supervised binary decisions can be approached using multiple methods. For a brief history, we can cite multivariate discriminant analysis and logistic regressions. To break free from traditional statistical assumptions, CART (Classification and Regression Trees) and machine learning methods are currently the most widely used methods in analysis processes, including in property and casualty insurance, which is gradually abandoning GLS methods for calculating risk premiums.

2.1. The C.A.R.T. method

The algorithm begins by selecting the explanatory variable which, thanks to its characteristics, best divides the population into two groups by maximizing inter-group variance.

The two groups created are called “nodes.” The operation is repeated until there is only one individual per group or according to a stopping criterion to be defined, which allows the final nodes, called leaves, to be obtained.

The second step proposes to minimize a function that takes into account the mean square error and the number of leaves. This function optimizes the complexity level of the tree in order to avoid overfitting. Overfitting would be to make each case a leaf. The result is an optimal tree.

The CART method is highly dependent on the order of the variables chosen and the variables chosen to build the predictive model, hence the importance of correctly performing the variable selection part. This limitation can be rectified by boosting or bagging techniques offered by machine learning.

The enormous advantage of CART methods is that they make decisions understandable through simple nodes. A decision or set of decisions is obtained by successively passing through a series of inequalities of explanatory variables (blue rectangles in Table 2 below). A leaf is correct when a majority of decision types are obtained (dominance of yes or no or green and red circles in Table 2).

2.2. Machine learning

The machine learning method offers global or ensemble learning based on the bagging technique. The objective is to train several models in order to propose a final model that combines their outputs. Bagging creates several subsets of learning data by random sampling with replacement. Bagging improves the stability and accuracy of predictions compared to a model obtained from a learning algorithm. It helps reduce entropy and avoid overfitting.

Technically, for each split, we no longer look for the best split among all explanatory variables (n), but rather the best split for p explanatory variables randomly drawn from n. This double “randomization” was introduced by L. BREIMAN [1]. The number of trees in the forest grows with the number of variables.

Despite all the subtleties of parameterization, models cannot escape overfitting. We will return to this subject later in this article and in the implementation of a set of algorithms. The more complex the algorithms, the greater the associated operational risk. The more complex the process, the lower the traceability of the explanatory variables. These methods, which are generally more complex than statistical methods, are often likened to black boxes due to the subtlety of the algorithms and their settings. Machine learning is confronted with the logic of interpretation and sharing of results. Finally, it must be recognized that machine learning is simpler to implement than the “deep learning” presented in Table 3, with variables that are not necessarily explanatory. Consequently, machine

learning must be used in conjunction with traditional approaches to refine hypotheses and choices. Table 2 summarizes the different approaches between C.A.R.T. (decision tree with inequalities) and random forest (ensemble learning of several models).

Table 2: C.A.R.T. and Random Forest[2]

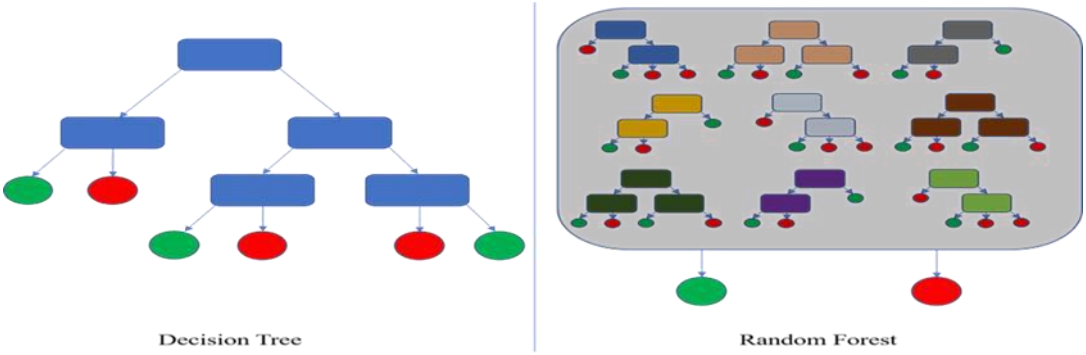
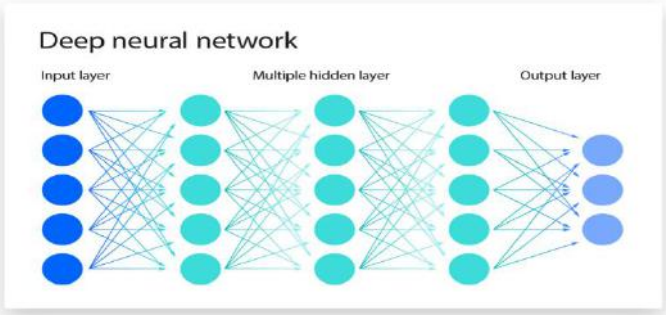


Table 3: Deep learning [3]



3. Methodological choices for implementing machine learning and its results

Creating the database was a long and complex process due to the multitude of information sources and databases.

A technical and analytical audit was necessary to identify the right databases and variables and to create a replicable database with the necessary historical data.

[2] https://www.pericles-group.com/
Machine Learning : Du GLM à l'arbre de CART en passant par le Random Forest
A Guide to Random Forest in Machine Learning novembre 2023

[3] https://www.ibm.com/think/topics/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks

The choice of technology was also a topic of discussion within the bank's management. For cost reasons, standard methods available in R or Python libraries were chosen. The standard methods for machine learning are the S.V.M. (support vector machine) and K.N.N. (K Nearest Neighbor) methods.

Numerous studies show the differences in performance between different machine learning techniques. One example is the article by Danilo Bzdok, Martin Krzywinski, and Naomi Altman [3].

3.1. K.N.N.

K.N.N. is a simple and highly effective supervised machine learning algorithm. It belongs to the family of non-parametric algorithms based on the similarity of input data points. K.N.N. essentially makes predictions based on the similarity of data points in the sampling space. The performance of KNN is essentially based on the choice of K. KNN works by memorizing the entire training data set. When a new data point is given for prediction, KNN examines the closest data points in the training set based on a specified distance metric (usually Euclidean distance). For classification, it assigns the majority class among the k nearest neighbors to the new data point. For regression, it predicts the average or weighted average of the target values of the k nearest neighbors.

- Advantages of K Nearest Neighbor (KNN)

Its implementation is simple. KNN is easy to understand and implement, making it suitable for rapid prototyping.

KNN is a learning algorithm that does not require a training phase. The model is built during the prediction phase.

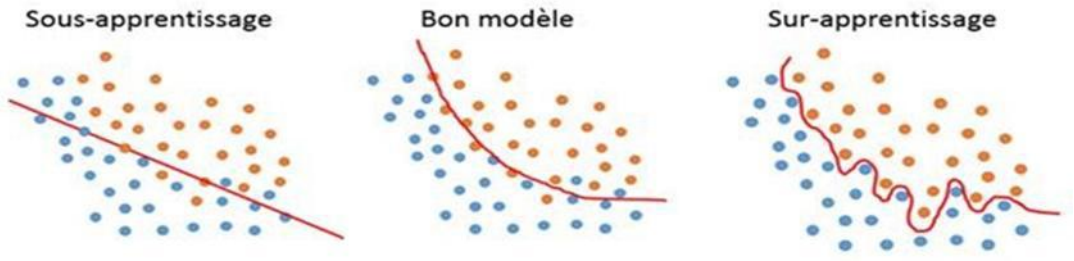
- Disadvantages of KNN

The main disadvantage is computational complexity. This increases with the size of the dataset. The computation required to find the nearest neighbors increases, resulting in higher computational costs. This method is also sensitive to outliers, significantly affecting the distances between points and, consequently, the predictions.

3.2. The S.V.M. method

S.V.M. finds the optimal hyperplane that maximizes the margin between the data points closest to the opposing classes. The margin is the distance between the hyperplane and the closest samples. These are called support vectors. The S.V.M. algorithm is widely used in machine learning because it can perform both linear and nonlinear classifications. When the data are not linearly separable, kernel functions are proposed to transform the data into a higher-dimensional space to allow for linear separation. This application of kernel functions is known as the “kernel trick.” The SVM method allows for linear classifications, which are often unsuitable, and nonlinear classifications (polynomial, etc.).

Table 4



To summarize, we can present a matrix of the strengths and weaknesses of the method.

- Advantages of the support vector machine (S.V.M.):

Efficiency in large spaces. SVM works well in high-dimensional spaces, making it suitable for tasks with a large number of features.

It is robust to overfitting. SVM has regularization parameters that help avoid overfitting while offering nonlinear models to separate classification spaces.

It is a global optimization method that requires a convex problem. If this is the case, it guarantees that the solution found is the global optimum. Effective in nonlinear data: with the use of kernel functions, SVM can handle nonlinear relationships between entities.

- Disadvantages of support vector machines (SVM):

The global optimum leads to complex calculations when dealing with large data sets. It is memory-intensive: particularly when dealing with large data sets, as the algorithm must store all support vectors. The method is sensitive to noise: S.V.M. is sensitive to noisy data, and outliers in the training set can have a significant impact on performance.

Selecting an appropriate kernel and adjusting it with parameters can be difficult, and the performance of the S.V.M. model is sensitive to these choices.

3.3. Methodological trade-offs

We chose a combination of SVM and CART. These hybrid methods are found in the scientific literature [9].

The methods as presented are all imperfect. The disadvantage of machine learning is that it generally partially escapes the explanatory processes of decision-making and remains sensitive to the quality of the training data.

In our research and implementation, we now have the essential data and variables thanks to hypothesis testing. We used CART to understand the decision-making process and identify anomalies in the initial database. Positive responses in negative response configurations raised serious questions. Identifying these cases allowed us to better understand the cases generating significant noise.

Specifically, we found cases where the decision cannot be understood without taking missing variables into account.

One example that can be cited is cases where overdrafts are accepted even though the customer has all the attributes of a bad customer (lack of assets, risk, low profitability for the bank). After reviewing the customer's file, we found that the bank authorized the overdraft thanks to a high-quality legal guarantee from the company's parent company. Without this information, the decision is incomprehensible and even dangerous to implement. There are two solutions. The first is to integrate this missing variable (guarantee and its quality). The second solution is to prune the information identified as noisy from the learning process. The bank chose the second solution because it was unable to reliably construct this computer variable. The data identified and removed from the learning base represented 3.5% of the original database.

We then used the SVM method on the refined data or decisions, a choice based on the quality of its global optimization method.

In addition, SVM is suited to our problem of two distinct classes (acceptances and refusals of overdrafts).

3.4 Results

There are multiple criteria for evaluating the statistical quality of the proposed ML model. The most commonly used is, of course, the confusion matrix, which identifies correct answers and others, or in other words, the risks of primary and secondary errors. We note that our database contains 432,966 decisions with 373,389 overdraft authorizations and 59,577 refusals spread across seven customer segments. We have, of course, created seven models that differ in both their composition and their performance.

In our study, we have four cases:

We have correctly modeled overdraft authorizations (true positive).

Correctly modeled overdraft refusals (true negative).

Overdraft authorizations not accepted by the model (false positive).

Overdraft refusals accepted by the models (false negatives)

The information for all segments is presented in Table 5.

It is clear that false positives and false negatives pose problems of acceptability for bank managers. False negatives wrongly authorize overdrafts. For this reason, the bank limits the thresholds for automated overdraft acceptance. False positives are also problematic because the account may be blocked. The latter case is not problematic in the sense that the bank retains the final decision on this block by manually triggering a payment stop. Overall, the models performed well in terms of the overall performance criterion of “accuracy” (true positive + true negative) / total. It should be noted that false positives are low, with 2,538 cases out of a total of 59,577 rejections (4.26% of cases).

(316,772 + 57,039)/432,966 = 87.10%

Table 5: Results – confusion matrix

	Modeling yes	Modeling no
Acceptance	316 772 true positive	56 617 False negative
Rejection	2 538 False positive	57 039 true negative

Table 6: Performance by segment

Customer segment	S1	S2	S3	S4	S5	S6	S7
« accuracy »	92.40	96.78	79.73	27.61	88.13	57.57	78.36

We found that the model does not work well when dealing with high entropy in two customer segments (S4 and S6). These two segments concern high-end retail customers, who are fewer in number than businesses and traditional individual customers. The entropy of S4 and S6 is linked to the complexity and uniqueness of high-end customer situations. Table 6 shows the overall performance for the seven segments.

The results show that automation has been effective in segments S1, S2, S3, S5, and S7, with an overall weighted performance of over 90%.

However, the bank has limited automation to amounts between €500 and €1,000, depending on the customer segment. This study shows the advantages and limitations of this automation. The advantages are clearly controlled automation in certain customer segments, with considerable time savings for bank employees. AI is a source of productivity gains but must remain a controlled process.

AI also generates development and maintenance costs that remain significant.

Sources of improvement remain in the refinement of databases and methodological advances that can be made in the field of machine learning. Here, the problem is the asymmetry of decisions between approvals and rejections. This asymmetry can limit the performance of traditional methods such as SVM.

4. A new method created specifically to address the methodological issue: the lack of symmetry in the number of decisions

The data in our field of study is obviously asymmetrical, with many more overdraft authorizations than refusals. This poses a challenge in terms of obtaining robust and reliable algorithms. A lack of reliability and robustness can have serious legal consequences for banks.

To address these decision asymmetries, the most recent approaches propose either methodological combinations as presented above (C.A.R.T. – machine learning – logistic regression, etc.) [7, 12, 13] or methodological advances in machine learning [15, 17].

4.1 Presentation of the DC programming and D.C.A. function

The SVM method offers global optimization. As part of a doctoral research program [15, 17], we adapted the method by proposing the DC programming and D.C.A. method. This method has proven its advantage in terms of results and computation time in many complex optimization problems [14]. The complexity of optimization methods often lies in the non-convex nature of the problem. The DCA method solves the convexity problem by transforming the initial function into two differentiated convex functions.

The formulation is as follows

$$\inf\{f(w) = G(w) - H(w) : w \in \mathrm{R^p}\}, (\mathrm{P_{dc}})$$

In the given context, we have the convex functions G and H ∈ Γ0 (Rp), which are the set of proper lower semi-continuous convex functions of a set Rp to R∪ {+∞}. These functions are called CC functions, where G−H represents a DC decomposition of the function f, G and H being the DC components. A convex constraint w ∈ C can be incorporated into the objective function of (Pdc).

4.2. Cost-sensitive weighted sampling based on DCA (CSB – DCA)

The goal of machine learning is to train several models or combinations of variables in order to propose a final model that combines their outputs. Bagging creates several subsets of training data by random sampling with replacement.

When the sample is unbalanced, the method can be disrupted by the preponderance of one decision. Our study does indeed contain many more favorable responses than negative responses to overdrafts. Cost-sensitive learning can correct this asymmetry. The method consists of assigning a greater weight to the minority class. In practice, the model considers that correctly classifying a decision in the minority class (in this case, rejection) is more important than correctly classifying a decision in the majority class (acceptance). This is why the S.V.M. and C.A.R.T. techniques previously implemented had the disadvantage of creating a significant imbalance between false positives and false negatives.

As shown in Table 2, each formalized response is the result of a weighted aggregation of different models. We can formalize this as follows:

$$f(x) = \sum_{j=1}^{n} w_j f_j\ (x)$$

Where f(x) is the aggregation of the models and w is the weight of each one.

Determining the appropriate weights is crucial for effective model performance. In the standard bagging scheme, each model is weighted identically. Breiman [2] proposes logically overweighing the models that offer the most popular choices.

Given a training dataset {(x_i, y_i)} to m, where m is the number of samples in the training dataset. Each sample is associated with a label y_i (1 or -1). The output of the aggregated model for the i-th data point is denoted {\hat{y}}_i. The objective is to minimize the prediction error, which gives rise to an optimization problem that will be solved here using the C.S.B.-D.C.A. method (cost sensitive based D.C.A.):

$$min\ k(w)_{w\epsilon R^n} = \frac{1}{m}\sum_{i=1}^{m} l(y_i, \hat{y}_i) = \frac{1}{m}\sum_{i=1}^{m} l(y_{i,} \sum_{j=1}^{n} w_j\ f_j\ (x_i))$$

$$w_j \geq 0, \forall j = 1, \ldots, n$$

where l is a loss that measures the difference between the predicted values and the actual values.

4.3. Comparative results with the S2 customer segment

A performance comparison was carried out between the two approaches, C.A.R.T. – S.V.M. and C.S.B. – D.C.A., using the usual criteria of overall performance "accuracy," F-score, Gmean, and AUC[4] on the S2 customer segment.

Table 7: Comparison of performance for the S2 customer segment

	Accuracy	Fscore	G mean	AUC
CART - SVM	97.20	98.28	98.57	98.56
CSB - DCA	99.98	99.75	99.78	99.78

The CSB-DCA model shows a slight improvement in convergence criteria and superiority over the CART-SVM combination. This method, which is adapted to response asymmetries, makes it possible to reduce false positives.

5. Conclusion and further research

The implementation of AI in a bank highlights the various dimensions that need to be addressed in research in this field. The first dimension is economic and managerial, profoundly changing the organization of work within banks and, more generally, within companies. AI is a source of productivity and resource reallocation. The second challenge is to understand the highly complex information systems that must serve as the basis for machine learning. This complexity leads us to favor hybrid methodologies, in this case combining the C.A.R.T. and S.V.M. methods to improve and understand the learning bases. However, this approach does not abandon upstream data analysis methods to improve understanding of decision-making processes.

The implementation of this approach is generally satisfactory, as it has made it possible to automate most decisions to exceed bank overdrafts, except for high-end customer segments, which are characterized by very high entropy.

The third challenge is technological. The bank has opted for Python-type open source libraries to install the algorithms for reasons of cost and comparability. Nevertheless, methodological improvements exist in machine learning when the responses to be modeled are asymmetric. Our C.S.B. – D.C.A. proposal improves the quality of convergence with observed reality. This method will be tested on broader customer segments and on its ability to withstand the introduction of random variables into the database.

[4] Fscore = true positive / (true positive + false positive).
AUC is a measure of the model's ability to distinguish between positive and negative classes. The ROC (Receiver Operating Characteristics) curve is a graphical representation of the model's performance, plotting the true positive rate (TPR) against the false positive rate (FPR) for different threshold parameters. To calculate the AUC in Matlab, we use the trapezoidal rule in the "Trapz" function. The AUC value ranges from 0 to 1, where a higher score indicates a better model.
G-mean is the geometric mean of the true positive response rates
true positive true positive + false negative × true positive true positive + false positive

$$\sqrt{\frac{true\ positif}{true\ positif\ +\ false\ negatif} \times \frac{true\ positif}{true\ positf\ +\ false\ positif}}$$

Références

1. Breiman, L. : random forests Published: October 2001 Volume 45, pages 5–32, (2001)
2. Breiman, L. : Bagging predictors. Machine learning 24(2), 123–140 (1996)
3. Bzdok, D., Krzywinski, M., Altman, N. : Machine learning: Supervised methods, SVM and kNN. Nature Methods, pp.1-6. (2018)
4. Chang, Y.C., Chang, K.H., Wu, G.J.: Application of extreme gradient boosting, trees in the construction of credit risk assessment models for financial institutions.
   Published in Applied Soft Computing 1 December (2018)
5. Damel, P. : Les produits structurés bancaires et le Contrôle de Gestion bancaire : une approche comparative utilisant les taux de marché de référence, communication avec publication dans les actes du colloque, 22ième congrès AFC (Mai 2001)
6. Damel, P. : L'apport des méthodes de « replicating portfolio » ou portefeuille répliqué en A.L.M. : méthode contrat par contrat ou par la valeur optimale », Banque et Marchés Mars-Avril (2001)
7. Kim, M.J., Kang, D.K., Kim, H.B.: Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction.
   Expert Systems with Applications 42(3), 1074–1082 (2015)
8. Hafez, IY. : A systematic review of AI-enhanced techniques in credit card fraud detection Journal of Big Data, SpringerOpen. (2025)
9. Komal Goyal, Megha Garg Shruti Malik Adoption of artificial intelligence-based credit risk assessment and fraud detection in the banking services: a hybrid approach (SEM-ANN) Future Business Journal volume 11, Article number: 44 (2025)
10. Guan, C. : Responsible Credit Risk Assessment with Machine Learning and Knowledge Acquisition Springer article "Human-Centric Intelligent Systems » SpringerLink (2023)
11. Lennart, H., V., Damásio, B. : Machine learning in banking risk management: Mapping decade of evolution. International Journal of Information Management Data Insights
    Volume 5, Issue 1, June 2025, 100324. (2025)
12. Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M.S., Zeineddine, H.: An experimental study with imbalanced classification approaches for credit card fraud
    detection. IEEE Access 7, 93010–93022 (2019)
13. Veganzones, D., S´everin, E.: An investigation of bankruptcy prediction in imbalanced
    datasets. Decision Support Systems 112, 111–124 (2018)
14. Pham Dinh, T., Le Thi, H.A.: Convex analysis approach to dc programming: theory,
    algorithms and applications. Acta Math. Vietnam 22(1), 289–355 (1997)
15. Pham Van Tuan (co-supervision Hoai An Lethi et Damel Pascal). "New machine learning techniques in financial decision making". Thesis defended in November 2023, at University of Lorraine (2023)
16. Roy, J.K. : Machine Learning and Artificial Intelligence Method for FinTech Credit Scoring and Risk Management : A Systematic Literature Review. International Journal of Business Analytics Volume 11 • Issue 1 • January-December (2024)
17. Van Tuan Pham, Hoai An Le Thi, Hoang Phuc Hau Luu, Damel Pascal: DCA-Based Weighted Bagging: A New Ensemble Learning Approach. Published ACIIDS (2) (2023)

18. Habib Zouaoui, H. : Credit card fraud detection and risk management strategies: A deep learning-based approach for EU banks. Research Papers in Economics and Finance Vol. 9 No. 1 (2025)