# Towards a Machine Learning-based Model for Corporate Loan Default Prediction

Imane RHZIOUAL BERRADA, Fatimazahra BARRAMOU, Omar BACHIR ALAMI
Laboratory of Systems Engineering, Hassania School of Public Works, Casablanca, Morocco

***Abstract*—As the core business of the banking system is to lend money and then get it back, loan default is one of the most crucial issues for commercial banks. With data analysis and artificial intelligence, extracting valuable information from historical data, to lower their losses, banks would be able to classify their customers and predict the probability of credit repayment instead of relying on traditional methods. As most actual research is focused on individuals' loans, the novelty of the present paper is to treat corporate loans. Its main objective is to propose a model to address the problem using selected machine learning algorithms to classify companies into two classes to be able to predict loan defaulters. This paper delves into the Corporate Loan Default Prediction Model (CLD PM), which is designed to forecast loan defaults in corporations. The model is grounded in the CRISP-DM process, commencing with comprehending corporate requirements and implementing classification techniques. The data acquisition and preparation phase are critical in testing the selected algorithms, which involve Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, XGBoost, and Adaboost. The model's efficacy is assessed using various metrics, namely Accuracy, Precision, Recall, F1 score, and AUC. Subsequently, the model is scrutinized using an actual dataset of loans for Moroccan real estate firms. The findings reveal that the Random Forest and XGBoost algorithms outperformed the others, with every metric surpassing 90%. This was accomplished by utilizing SMOTE as an oversampling method, given the dataset's imbalance. Furthermore, when concentrating on financial statements, selecting the five most significant financial ratios and the company's age, Random Forest was adept at predicting defaulters with good results: accuracy of 90%, precision of 75%, recall of 50%, F1 score of 60% and AUC of 77%.**

***Keywords*—Loan default; prediction; artificial intelligence; data analysis; machine learning; companies; corporate; real estate; bank**

## I. Introduction

Banks worldwide need to secure their loans and minimize defaults to maintain healthy financial results. To achieve this, they use various risk rating methods based on traditional approaches or more recently, innovative ones that incorporate artificial intelligence. It is thus important to classify customers to predict their worthiness (Ability to pay back the loan) [1] before loans approval.

The banking industry deals with an enormous amount of data on a daily basis. As a result, financial institutions need to rely on the outcome of this data to strengthen their risk strategy. Credit scoring has become a competitive advantage for these institutions in order to optimize their profits, as reported by [2]. There are several steps that they should follow to achieve this objective. To start with, data analysis should be a central issue in the decision-making process of approving or rejecting a loan.

Retail banking (individuals' loans) has been studied in different research works and recent articles thanks to open and various available datasets [3] [4] [5]. Many literature reviews are also available for personal credit scoring [6] [7].

Moreover, commercial banks can experience significant losses due to default payments on loans, particularly in cases where large amounts are involved, such as financing investment projects. However, there is limited research on this topic [8] [9], with most studies focusing on personal loans rather than investment loans for companies. Therefore, this research will specifically address the issue of corporate loans.

Companies have been affected by various recent crises worldwide, including in Morocco. It is widely recognized that investment projects are essential for the success of every economy and are crucial for all industries. To this end, banks play a critical role in approving loans for companies seeking financing for development. As explained by [10] studying the drivers of default, central banks and governments have a concern to ensure balanced growth in the market.

While studying actual research, no detailed model was found with description from the beginning of the project till its testing phase and implementation. A guided step by step model to follow and apply is needed for financial institutions and researchers.

In this article, the Corporate Loan Default Prediction Model (CLD PM) is presented with its detailed roadmap and application results to be used by researchers and banks to predict losses and avoid risky loan distribution for companies. It is a model based on CRISP-DM for which the steps are detailed. For that, the related work concerning the process is presented. Then, the algorithms and metrics used are highlighted. After that the model with the chosen machine learning algorithms and evaluation metrics is exposed, to end up with the results for the application on a real-world dataset of real estate development loans for Moroccan companies. In conclusion, the test results are detailed as well as the limitations, next steps, and perspectives to work on.

The novelty of this article is that it proposes a comprehensive approach for investment loans offered to companies. The default on these loans can be attributed to various factors such as the financial health, history, behavior, and qualitative data of the company. The implementation phase

is based on a thorough analysis of financial statements and ratios that differentiate corporates.

## II. Related Work

Several articles discussing data mining and data science methods have been reviewed, and the most relevant ones have been selected for presentation here. Additionally, a comparison of different machine learning algorithms for classification has been conducted based on a previous review, and the algorithms of interest for testing have been narrowed down. It will be presented in the following section with a specific focus on the problem. The review of articles dealing with data mining and data science methodologies has resulted in the selection of the most relevant ones for presentation. Furthermore, a comparison of machine learning algorithms for classification has been conducted based on the previous review to limit the interest to the algorithms to test [11].

The related work will be presented in the following section with a specific point of view for the problem.

### A. Related Work: Process

After reviewing various historical process models, it seems that CRISP-DM (Cross-Industry Standard Process for Data Mining) is an interesting process model, which inspires the theoretical approach before delving into applications and testing (see Fig. 1).

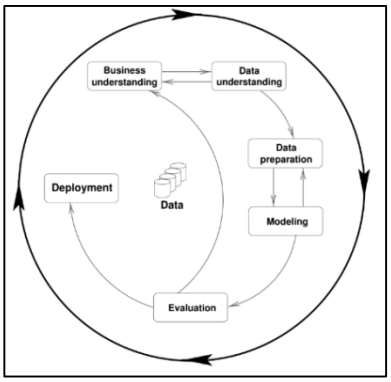


Fig. 1. CRISP DM [15].

Indeed, according to S. Saltz et al. in 2016 through their article [12], to hold a successful project, there is a need for a process, key process attributes, and effective team communication. This literature review for many thousands of conferences and articles highlighted two well-known models for Data Mining: CRISP-DM and SEMMA (Sample, Explore, Modify, Model, and Assess) confirming that they might be not appropriate for BD (Big Data) projects. The article demonstrates that Agile methodologies have more advantages than waterfall methodologies.

CRISP-DM has been established in the middle of the nineties based on previous models and is the most well-known and used process. It relies on six steps: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [8] [13]. SEMMA was developed by the SAS institute and is the second most popular methodology [14].

These methods have similarities and more research focus on new methodologies based on these.

In the continuity of the previous work of S. Saltz et al, for [16], an experiment was held comparing four different methodologies with four different teams holding projects. Evaluated by independent experts and through stakeholders' surveys, Agile Kanban (based on the principle of moving quickly and easily by focusing on small parts of the project) and CRISP-DM outperformed.

Then, different developed methodologies are found based on the previous cited and others such as DMME [17] that is an extension to the CRISP-DM adding some adjustments to adapt the methodology taking into account engineers' points of view. This method focuses on data collection and acquisition methods, technical understanding, and workflow monitoring while the projects in the run for more effectiveness.

F. Martinez et al. [18] studied the evolution twenty years after CRISP-DM was introduced. They present the move from the Data mining process and its first discovery to Data Science trajectories. An interesting figure presents different models and methodologies derived mainly from KDD and CRISP-DM. A diagram has been proposed to include all the activities identified in a data science project. This will help define different trajectories for a customized Data Science Trajectories model, called DST, for each project. The diagram, which is shown in Fig. 2, includes all activities from data management, CRISP-DM and exploratory. It is possible to have one or many trajectories for each project, which can include any of these activities.

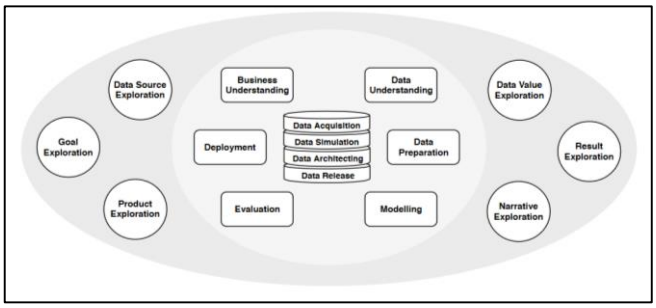


Fig. 2. DST MAP [18].

### B. Related Work: Algorithms and Metrics

The problem concerning loan default prediction is a classification issue needing appropriate algorithms to perform and compare. As in a previous work, the different possible approaches were studied and the most effectively used algorithms were identified [11], the algorithms to be tested in the implementation phase are the following: Logistic regression, Decision tree, Random Forest, Support Vector Machine, Xgboost and Adaboost. Moreover, other recent articles highlight interesting results with these algorithms [19] [20]. In the following, a brief description of each is presented.

Logistic regression (LR) is a statistical method used to predict if there is a certain value as an outcome of a probability. Authors of [21] prove the outperformance of logistic regression according to ROC.

Decision tree (DT) is a graphical representation of possible solutions to a decision based on certain conditions. In their article, authors of [22] present a comparison between decision tree and random forest in which random forest outperforms.

Random forest (RF) is an ensemble algorithm, bagging the decision tree. It creates multiple decision trees in the training process. Authors of [22] compare the performance of decision trees and random forest and random forest outperformed. Moreover, the article of [23] compares the performance of ensemble algorithms concluding that ensemble algorithms have better results. This algorithm performs well even with thousands of variables according to the authors of the article.

Support Vector Machine (SVM) is a large-margin classifier that tries to find the maximum margin separating the dataset into two categories. Moreover, [24] compared random forest to SVM and found that random forest was quick and had more simplicity whereas SVM had better accuracy. In their article, authors of [1] performed a two steps testing applying first random forest then SVM for good performance.

eXtreme Gradient Boosting (XgBoost) is an ensemble learning method using a collection of weak learners (decision trees) to have strong predictions. It was combined with Lightgbm to propose a hybrid model by Z. Song [25] outperforming for fraud detection.

Adaptive Boosting (AdaBoost) is a general boosting algorithm ensemble learning combining individual weak learners with sequential adjusted weights to improve accuracy. According to authors of [26], AdaBoost reaches the highest performance.

Concerning evaluation metrics, there are several used to evaluate machine learning algorithms' performance. In the following, the most important and commonly used ones for classification problems are presented [11]:

- Accuracy: Proportion of true among total
- Precision: Proportion of the predicted positive cases that are correct
- Recall or sensitivity: Proportion of positive cases that are correctly identified
- F1 score: Weighted harmonic mean of precision and recall
- AUC (Area Under the ROC Curve): Probability that a random positive is positioned to the right of a random negative ROC plots

Many actual studies use all these evaluation metrics and others to choose the best one. Authors of [27] choose accuracy to conclude that random forest delivered the best results compared to other algorithms.

L Zhang et al. presented in their article [4] a metric according to profit for peer-to-peer lending as accuracy can be non-sufficient. But most of the tested comparisons lead to accuracy as a key performance indicator.

A very interesting literature review held in [28] analyses and figures out the metrics tested and used by different articles studied. This article also tackles a specific issue concerning the evolution of credit scoring evaluation and research studies still lacking today.

### *C. Related Work: Companies Loan Default Prediction*

Early studies started in 1966 about bankruptcy prediction with statistical methods based on previous available data about the companies. In 1999, with [29], discriminant analysis is used and performed well for prediction. Then, in 2005 with [30], Back propagation neural networks and SVM were used for small datasets to predict companies bankruptcy which is an advanced level of default. Indeed, a company that goes on bankruptcy is subsequently a defaulter regarding its creditors.

Authors of [26] presented the literature review concerning corporate credit risk as well as consumer and P2P. They highlighted that companies' loans are the most important ones for banks. Some of previous works were presented including those held from the credit crisis in 2007 [31] , 2012 [32] and 2014 [33]. These studies explored different SVM applications and variants to confirm their good performance.

For [8], researchers studied companies loan default prediction and applied machine learning algorithms for classification to demonstrate the superiority of Random forest. Moreover, [9] examined credit risk assessment and confirmed that SVM have good accuracy while applied to a company's dataset limited to three features. They also studied the impact of the company daily income to its credit score.

In the same register, authors of [34] proposed a combined model for SME (Small and Medium Entreprises) comparing the performance of separated SVM and combined and optimized with rough sets to identify key factors influencing credit risk by reducing classification indicators.

On another hand, [35] handles the problem facing industrial companies in India after the COVID crisis lowering their capacity to pay their loans and avoid bankruptcy focusing on some key predictor financial ratios. Financial data for companies can thus deliver hidden information and lead to default prediction.

Moreover, in [36], a comparison is held between statistical and machine learning classification. The conclusion leads to the added value of machine learning for datasets with few features. F. Azayite et al. [37] propose a hybrid model combining discriminant analysis, multilayer neural networks and self-organizing maps. They confirm then that a model performed with the appropriate data deliver better results.

Hyeongjun et al. presented a literature review [38] for companies loan prediction highlighting many limitations of the actual research and focusing on the importance of data governance and financial engineering to benefit from machine learning algorithms with good data preprocessing.

A great analysis was held by M Modina et al. [39] according to accounting data and credit indicators from private internal sources about previous loans history. Both indicators provide valuable information and predictive capabilities. The impact of each feature is studied. Moreover, authors raise the fact that the results could vary depending on the sector and location to explore for future work.

### D. Related Work: Imbalanced Datasets

Concerning imbalanced datasets for companies, [40] analyses SMOTE (Synthetic Minority Over-sampling Technique) and combined Weighted SMOTE with ensemble learning (random forest) in order to propose a solution to the cited problem for small business. Authors of [41] also used SMOTE to tackle the problem and reach good results.

Moreover, [42] tested different resampling methods and concluded that among oversampling, undersampling, both and SMOTE, SMOTE outperforms. [43]

## III. Corporate Loan Default Prediction Model

To accurately identify defaulters and non-defaulters in a business, a Corporate Loan Default Prediction Model (CLD PM) has been proposed based on previous research. This model takes into account the subject specificities, while focusing on the limitations of each approach. It uses the CRISP-DM methodology.

However, before modeling and evaluation, it is recommended that data preparation is thorough and appropriate. With different data preparation, different results are obtained. Using machine learning algorithms that have been successful in the past can help run tests and select the best algorithm for deployment. By prioritizing data preparation, it is possible to improve the accuracy of the predictions and make informed decisions for business.

The CLD PM is presented in Fig. 3 and detailed in the following. The model was validated with the use case dataset:

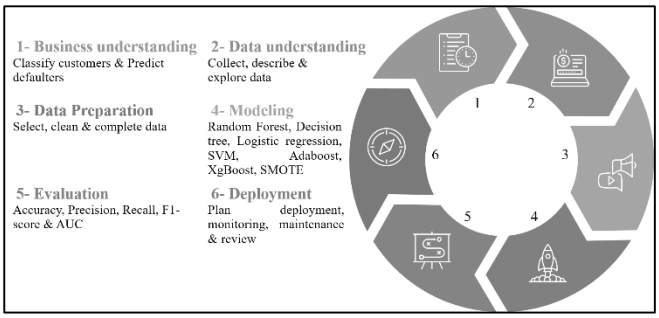


Fig. 3. Corporate Loan Default Prediction Model (CLD PM).

### A. Business / Problem Understanding

In this step, it is a must to understand the business, and describe the problem faced. Banks distribute loans to companies and individuals but the problem encountered is risk of default which might be minimized in the approval phase. It is a binary classification problem (« defaulters » and « non-defaulters » / « good » and « bad » payers).

Hence, it is necessary to define the purpose to reach in order to approve the model's results. The objective is to find a solution to the problem with machine learning and obtain good performance of classification.

### B. Data Understanding / Collection

At this step, there is normally no available dataset to perform the model on, there is a necessity to identify the needed data with its attributes for each record. In the following, according to business knowledge, a list of identified features is identified to distinguish different payers' profiles to classify them into "good" and "bad" payers:

- Data concerning the company (Identification, Activity, size based on the annual turnover, financial ratios & data from financial statements, experience, quality of management…)
- Data concerning banks' relationship (Transactional data, Credit score, other banks historical data…)
- Data concerning the loan characteristics (Loan type, release date, Default…)

After needed data identification, the acquisition process can be launched.

Unfortunately, most of the time, while facing the step of data acquisition, it seems complicated to collect the defined features even if existing in the bank's several separated systems and databases. The data acquisition goes through a long and complicated process. The combination of these data provides the dataset to deal with.

When the dataset is available, an important step is to understand its content and confirm that it is conform to what was requested. If the output doesn't fulfill the aimed dataset, a loop to the second step is needed for readjusting. It is a validation step before starting data analysis and AI modeling and testing.

If the output dataset is satisfying in terms of identified needed features, the description and visualization can be held with line charts, bar charts, heatmaps, and others before preprocessing.

### C. Data Preparation

This step plays a central role and has to be correctly held to strengthen the model's performance. It is the looping node and the crucial treatment in the model. In the following, the detailed steps are presented in order to perform machine learning algorithms:

- Feature selection with a business knowledge insight
- Conversion of categorical data to float
- Conversion of dates to years
- Reduce features dropping highly correlated ones
- Drop features with more than 50% missing values
- Complete missing values with the most frequent values

It is also possible to visualize a heatmap and select the most important features after performing Data preparation. Moreover, for loan default prediction, the datasets are imbalanced with a minority class of default. SMOTE is here the chosen technic to handle the problem.

### D. Modeling - Machine Learning Training

As previously cited, the following algorithms are performed:

- Logistic regression (LR)

- Decision tree (DT)
- Random Forest (RF)
- Support Vector Machine (SVM)
- eXtreme gradient Boosting (XgBoost)
- Adaptive Boosting (AdaBoost)

To train the prepared dataset, the dataset is split into a training set of 80% and 20% for testing as used for training classification machine learning algorithms [44].

### E. Evaluation

To appreciate the performance of each algorithm, the following metrics are tested with and without SMOTE:

- Accuracy
- Precision
- Recall or sensitivity
- F1 score
- AUC (Area Under the ROC Curve)

As long as the results of all the algorithms concerning all the metrics don't exceed a certain predefined value, a loop and readjustment of models parameters and preprocessing are performed. If only one algorithm performs well, it can be adopted for implementation.

## IV. Case Study

The case study to implement the previous model is a dataset of real estate companies from a commercial Moroccan bank.

Concerning the tools and the environment for implementation, the following are chosen:

- Integrated development environment: Jupiter Notebook
- Dataframe: Pandas
- Machine learning libraries: Scikit-learn, Pytorch, Matplotlib, seaborn, Numpy
- Programming languages: Python

To begin with, the problem at hand is to classify loan applicants for predicting defaults among companies. The objective is to achieve good metrics performance by utilizing the available data. The objective is to reach outstanding metrics with a target of more than 0.9 for all metrics.

For data identification, the meaningful features from the perspective of business experts are listed. Unfortunately, all the selected features weren't extracted. The available data collected from different sources and their combination is a fact to deal with. Data understanding and visualization are needed to validate and perform the rest of the model.

Before analyzing the dataset and visualizing it, features are defined. As there are 107 features, the most meaningful ones with expert's insight are below in Table I. An advanced analysis with the impact of each features and a classification of their importance according to each machine learning algorithm can be performed in further work analysis:

The dataset contains 396 records with 107 features for companies with loans released from 2015 to 2020. It contains companies' historical, qualitative, and financial data. It has 48 large companies and 348 Small and Medium-sized enterprises as shown in Fig. 4.

Concerning default, there are 50 defaulters and 346 non-defaults as shown in Fig. 5. It is an imbalanced dataset for which additional processing is needed.

Furthermore, among all the features, there is a correlation and some data with no economic sense for the present problem. For values, there are categorical data, dates, and missing values. To handle these issues, feature reduction is performed with multiple loopings to the test phase as the results needed to be meaningful and satisfying.

TABLE I. Description of Most Important Dataset Features

| Feature name | Brief description | Data type |
|---|---|---|
| Ref | A unique ID to loan application | Numeric |
| Annee | The year of loan approval | Date |
| Segment | The size of the company based on its Turnover (GE for Big companies & PME for Small & Medium companies) | Categorical |
| CATEGORIE JUR | The legal category of the company | Categorical |
| Anciennete entreprise | The age of the company | Numeric |
| Anciennete relation | Relationship age | Numeric |
| MAX NBR JOUR DEBITEUR | Maximum number of debtor days | Numeric |
| sum mcm net mad | Sum of credit movement | Numeric |
| Score crédit bureau | Credit score | Numeric |
| EBE CA | EBITDA (Earning before interest, tax, depreciation and amortization) on turnover | Numeric |
| TRES TB | Net cash on total assets (Liquidity ratio) | Numeric |
| FR FINAN EBE | Financial costs on EBITDA (Debt ratio) | Numeric |
| dettef kpropres | Financial debts on equity (Debt ratio) | Numeric |
| stk ca | Stock on turnover (Activity ratio) | Numeric |
| RN CA | Net profit on turnover (Profitability ratio) | Numeric |
| RN KP | Return on equity (Profitability ratio) | Numeric |

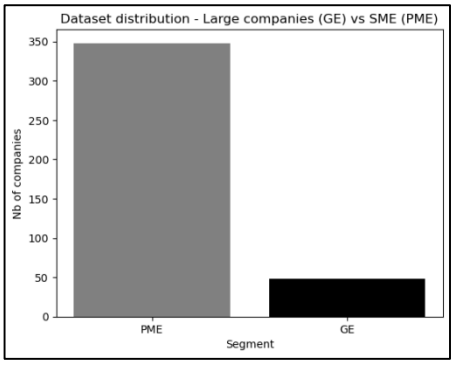


Fig. 4. Dataset distribution – Large companies and SME.

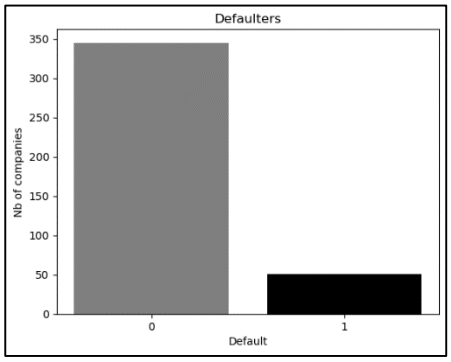


Fig. 5. Dataset distribution – Defaulters and non defaulters.

## V. Results and Discussion

The six algorithms are trained and tested the five metrics to obtain results adjusting hyper parameters to maximize the results. Table II and Fig. 6 illustrate the results of the maximized results.

The metrics measures are not satisfying even if accuracy and AUC can present good scores in some cases, precision, recall and F1 score underline bad performance as they are between 0 and 0.66. All the algorithms don't perform well. The identified origin of the problem is the imbalanced dataset. Default is only 13% of the dataset and the split of the dataset into the training of 80% and testing 20% lowers the probability of having a balanced dataset while testing on the small dataset of 396 records.

To tackle the problem of the small class imbalanced dataset, SMOTE was tested to resample the dataset. It is an oversampling technique that generates synthetic samples to resolve the problem of class minority and have a balanced distribution. It uses KNN and interpolating. Table III and Fig. 7 highlight the results of the test phase while using SMOTE.

TABLE II. Metrics of Tested Algorithms without SMOTE

| Algorithm / Metric | Accuracy | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|
| Decision tree | 0.8250 | 0.3750 | 0.2500 | 0.3000 | 0.64 |
| Random Forest | 0.8500 | 0.5000 | 0.0833 | 0.1429 | 0.89 |
| SVM | 0.8500 | 1.0000 | 0.0000 | 0.0000 | 0.64 |
| Logistic Regression | 0.8000 | 0.3000 | 0.2500 | 0.2727 | 0.59 |
| XGBoost | 0.8625 | 0.6667 | 0.1667 | 0.2667 | 0.88 |
| AdaBoost | 0.8500 | 0.5000 | 0.2500 | 0.3333 | 0.64 |

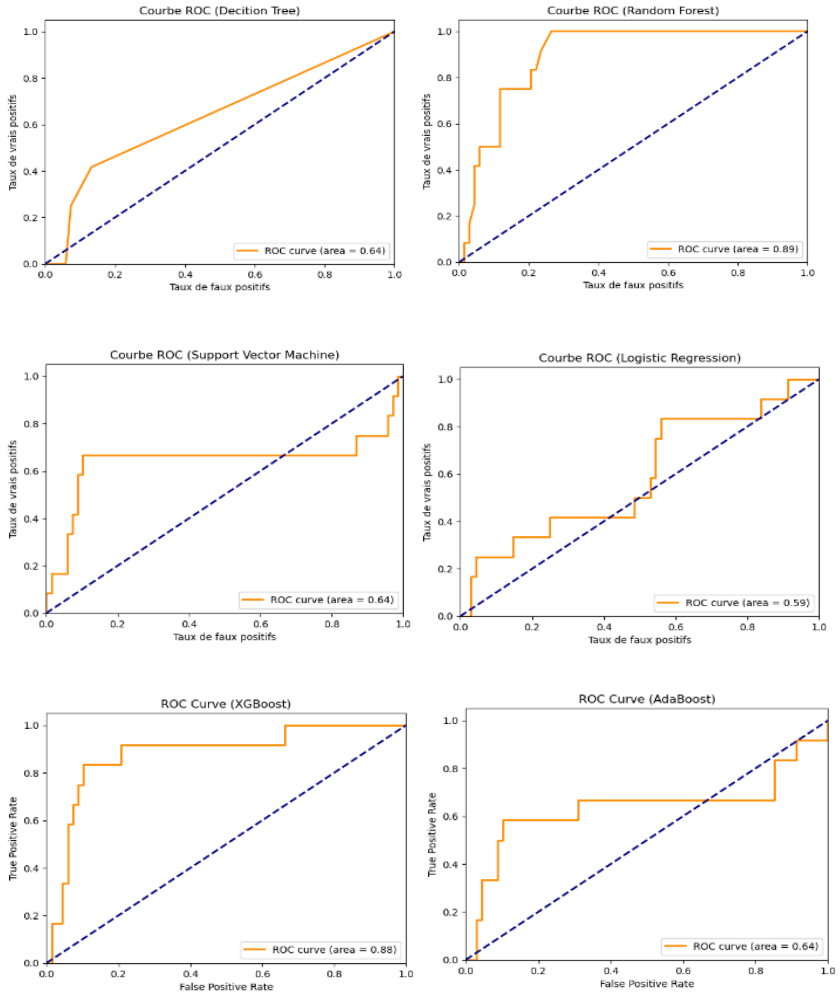


Fig. 6. ROC curve of the 6 algorithms without SMOTE.

TABLE III. METRICS OF TESTED ALGORITHMS WITH SMOTE

| Algorithm / Metric | Accuracy | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|
| Decision tree | 0.8188 | 0.7692 | 0.8333 | 0.8000 | 0.85 |
| Random Forest | 0.9420 | 0.9062 | 0.9667 | 0.9355 | 0.99 |
| SVM | 0.8875 | 0.6154 | 0.6667 | 0.6400 | 0.77 |
| Logistic Regression | 0.7536 | 0.7097 | 0.7333 | 0.7213 | 0.82 |
| XGBoost | 0.9420 | 0.9333 | 0.9333 | 0.9333 | 0.98 |
| AdaBoost | 0.9130 | 0.8636 | 0.9500 | 0.9048 | 0.97 |

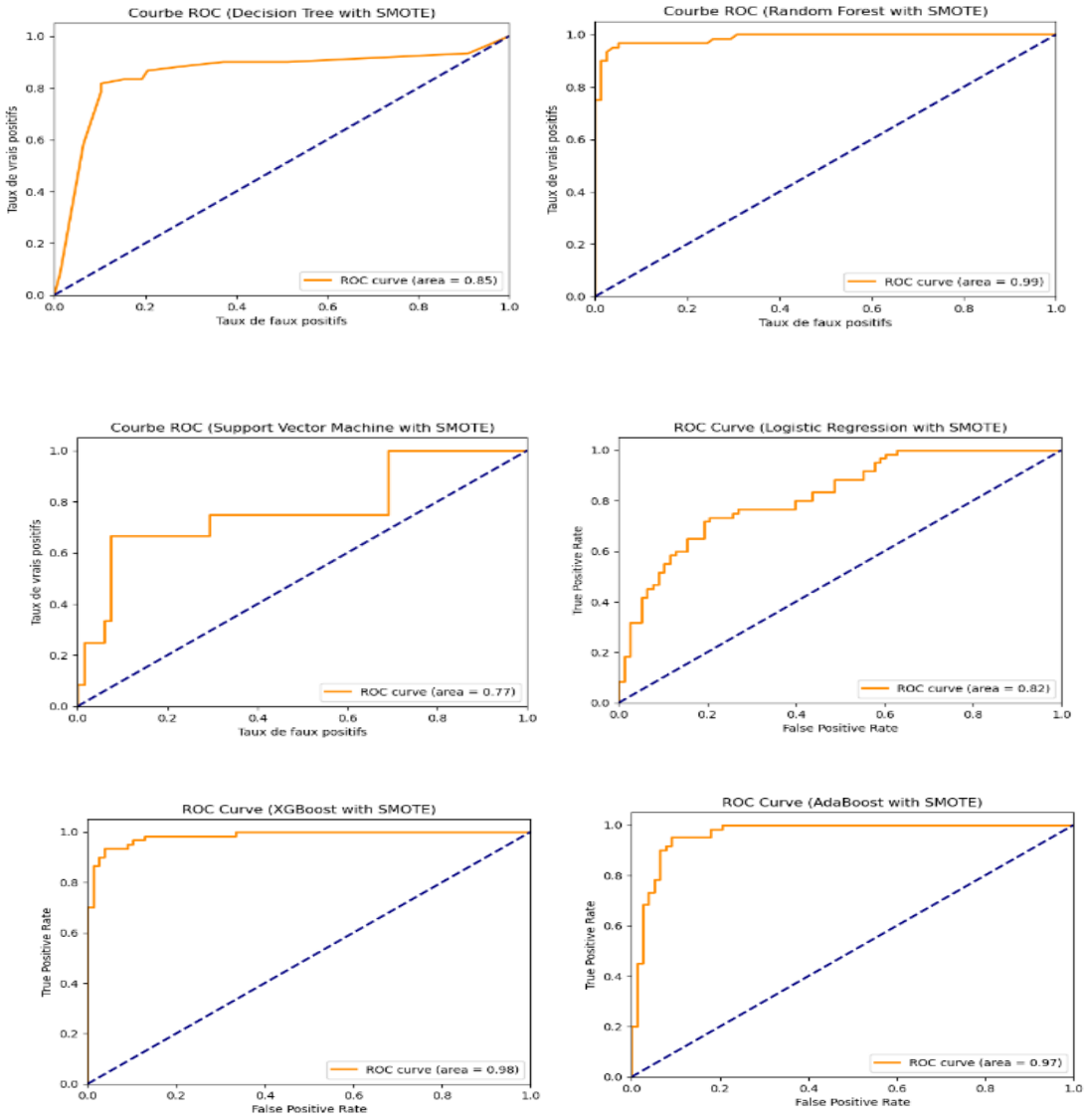


Fig. 7. ROC Curve of the 6 algorithms with SMOTE.

Based on the table and figures provided, it can be concluded that SMOTE enhances the performance of all the algorithms across all the metrics. The oversampling technique using SMOTE results in improved predictions, with fewer false positives and false negatives. The top-performing algorithms are Random Forest and XGBoost. As the approach is based on actual research results combined with knowledge concerning business problem, some external factors should disturb the expected results. The macroeconomic context, the Covid-19 crises, and the financial crises caused by the Russia / Ukraine conflict as well as the growth of inflation rate and loan interest rate consequently, will bias predictions' outcomes. There must be readjustment to perform. Future research can handle this issue. Indeed, historical data for some examples can make no sense before 2020 as some companies went into bankruptcy even if they had robust financial health before the COVID

crises. The sector in which the model is operated has its external factors to take into account.

In this context, as highlighted by [35], it is obvious that some financial ratios have significant impact on the prediction outcome. Indeed, a more advanced analysis allows to classify the most important features for each algorithm performed. With their combination, a set of five important features consisting of ratios from financial statements are identified:

- Liquidity ratio: Net cash on total assets (TRES TB)
- Debt ratio: Financial debts on equity (dettef kpropres)
- Profitability ratio - Return on equity: Net income on equity (RN KP)
- Profitability ratio – Margin: EBITDA on turnover (EBE CA)
- Activity ratio: Stock on turnover (stk ca)

When performing Random Forest with SMOTE on the dataset limited to these five ratios, without taking into account the other features concerning the company and its banking behavior, the model output allows to reach good performance for Random forest as presented below in Table IV:

Furthermore, if the age of the company is added to the five financial ratios, the model is strengthened as shown in Table V and can be applied to real estate companies.

TABLE IV. METRICS OF RANDOM FOREST WITH SMOTE – 5 FINANCIAL RATIOS

| Algorithm / Metric | Accuracy | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|
| Random Forest | 0.8500 | 0.500000 | 0.583333 | 0.538462 | 0.738971 |

TABLE V. METRICS OF RANDOM FOREST WITH SMOTE – 5 FINANCIAL RATIOS & THE COMPANY AGE

| Algorithm / Metric | Accuracy | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|
| Random Forest | 0.9000 | 0.750000 | 0.500000 | 0.600000 | 0.768995 |

As the use case considered has a small imbalanced dataset, a bigger dataset is needed to test and validate the model and the results. Moreover, future work can be held considering larger dataset with companies in different industries. The financial ratios to adopt for other industries might be different from real estate companies as this sector has specific accounting rules and midterm cycle projects development.

It would be also interesting to test the model proposed by [5] using a multi-classification method rather than a binary considering late payers not as defaulters but as a third class.

## VI. CONCLUSION AND PERSPECTIVES

In the present article, a model is proposed with its detailed steps for the loan default prediction using machine learning applied to Corporate Loan Default Prediction Model (CLD PM). With few founded research concerning this field and no proposed model detailing the process with all its components, algorithms and metrics, the present article offers an overview applied to a dataset of 396 loans for real estate companies.

Accuracy, precision, recall, F1 score and AUC for six different algorithms were compared. Moreover, SMOTE was used to conclude that for an imbalanced datasets, with data concerning the company, its financial statements and bank relationship, Random Forest, and XgBoost outperform. For five selected most important features (financial ratios) from different most important features of the algorithms performed with the age of the company, random forest with SMOTE can be applied.

For future work, additional data concerning the projects, their market and their location could also have an influence on the output and lead to different results.

## REFERENCES

[1] G. Roy et S. Urolagin, « Credit Risk Assessment Using Decision Tree and Support Vector Machine Based Data Analytics », in Creative Business and Social Innovations for a Sustainable Future, M. Mateev et P. Poutziouris, Éd., in Advances in Science, Technology & Innovation. Cham: Springer International Publishing, 2019, p. 79-84.

[2] M. Anand, A. Velu, et P. Whig, « Prediction of loan behaviour with machine learning models for secure banking », Journal of Computer Science and Engineering (JCSE), vol. 3, no 1, p. 1-13, 2022.

[3] X. Zhang et L. Yu, « Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods », Expert Systems with Applications, p. 121484, 2023.

[4] L. Zhang, J. Wang, et Z. Liu, « What should lenders be more concerned about? Developing a profit-driven loan default prediction model », Expert Systems with Applications, vol. 213, p. 118938, 2023.

[5] F. Alghamdi et N. Alkhamees, « DefBDet: An Intelligent Default Borrowers Detection Model », International Journal of Advanced Computer Science and Applications, vol. 14, no 7, 2023.

[6] J. A. Ogosi Auqui, J. Cano Chuqui, V. H. Guadalupe Mori, et D. H. Obando Pacheco, « Machine learning for personal credit evaluation: A systematic review », 2022.

[7] N. Suhadolnik, J. Ueyama, et S. Da Silva, « Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach », Journal of Risk and Financial Management, vol. 16, no 12, p. 496, 2023.

[8] C. Nejjar, M. Kaicer, S. E. Haimer, A. Idhmad, et L. Essairh, « Credit Risk Management in Microfinance: Application of Non-repayment Prediction Models », in International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD'2023), vol. 930, M. Ezziyyani, J. Kacprzyk, et V. E. Balas, Éd., in Lecture Notes in Networks and Systems, vol. 930. , Cham: Springer Nature Switzerland, 2024, p. 301-308.

[9] Z. Dai, Z. Yuchen, A. Li, et G. Qian, « The application of machine learning in bank credit rating prediction and risk assessment », in 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), mars 2021, p. 986-989.

[10] A. Saha, L. Hock Eam, et S. Goh Yeok, « Housing loan default in Malaysia: an analytical insight and policy implications », International Journal of Housing Markets and Analysis, vol. 16, no 2, p. 273-291, 2023.

[11] I. R. Berrada, F. Z. Barramou, et O. B. Alami, « A review of Artificial Intelligence approach for credit risk assessment », in 2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP), IEEE, 2022, p. 1-5.

[12] J. S. Saltz et I. Shamshurin, « Big data team process methodologies: A literature review and the identification of key factors for a project's success », in 2016 IEEE International Conference on Big Data (Big Data), IEEE, 2016, p. 2872-2879.

[13] C. Schröer, F. Kruse, et J. M. Gómez, « A systematic literature review on applying CRISP-DM process model », Procedia Computer Science, vol. 181, p. 526-534, 2021.

[14] A. Azevedo et M. F. Santos, « KDD, SEMMA and CRISP-DM: a parallel overview », IADS-DM, 2008.

[15] D. S. Putler et R. E. Krider, Customer and Business Analytics: Applied Data Mining for Business Decision Making Using R. CRC Press, 2012.

[16] J. Saltz et K. Crowston, « Comparing data science project management methodologies via a controlled experiment », 2017.

[17] H. Wiemer, L. Drowatzky, et S. Ihlenfeldt, « Data mining methodology for engineering applications (DMME)—A holistic extension to the CRISP-DM model », Applied Sciences, vol. 9, no 12, p. 2407, 2019.

[18] F. Martínez-Plumed et al., « CRISP-DM twenty years later: From data mining processes to data science trajectories », IEEE Transactions on Knowledge and Data Engineering, vol. 33, no 8, p. 3048-3061, 2019.

[19] S. Kumar et al., « Exploitation of Machine Learning Algorithms for Detecting Financial Crimes Based on Customers' Behavior », Sustainability, vol. 14, no 21, Art. no 21, janv. 2022.

[20] H. I. T. Aziz, A. Sohail, U. Aslam, et N. Batcha, « Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess (SEMMA) », Journal of Computational and Theoretical Nanoscience, vol. 16, p. 3489-3503, août 2019.

[21] P. Maheswari et C. V. Narayana, « Predictions of Loan Defaulter - A Data Science Perspective », in 2020 5th International Conference on Computing, Communication and Security (ICCCS), oct. 2020, p. 1-4.

[22] M. Madaan, A. Kumar, C. Keshri, R. Jain, et P. Nagrath, « Loan default prediction using decision trees and random forest: A comparative study », in IOP Conference Series: Materials Science and Engineering, IOP Publishing, 2021, p. 012042.

[23] Y. Li et W. Chen, « A comparative performance assessment of ensemble learning for credit scoring », Mathematics, vol. 8, no 10, p. 1756, 2020.

[24] G. Teles, J. J. P. C. Rodrigues, R. A. L. Rabêlo, et S. A. Kozlov, « Comparative study of support vector machines and random forests machine learning algorithms on credit operation », Software: Practice and Experience, vol. 51, no 12, p. 2492-2500, 2021, doi: 10.1002/spe.2842.

[25] Z. Song, « A Data Mining Based Fraud Detection Hybrid Algorithm in E-bank », in 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), juin 2020, p. 44-47. doi: 10.1109/ICBAIE49996.2020.00016.

[26] L. Lai, « Loan Default Prediction with Machine Learning Techniques », in 2020 International Conference on Computer Communication and Network Security (CCNS), août 2020, p. 5-9. doi: 10.1109/CCNS50731.2020.00009.

[27] V. Aithal et R. D. Jathanna, « Credit risk assessment using machine learning techniques », International Journal of Innovative Technology and Exploring Engineering, vol. 9, no 1, p. 3482-3486, 2019.

[28] A. Markov, Z. Seleznyova, et V. Lapshin, « Credit scoring methods: Latest trends and points to consider », The Journal of Finance and Data Science, vol. 8, p. 180-201, nov. 2022, doi: 10.1016/j.jfds.2022.07.002.

[29] Z. R. Yang, M. B. Platt, et H. D. Platt, « Probabilistic neural networks in bankruptcy prediction », Journal of business research, vol. 44, no 2, p. 67-74, 1999.

[30] K.-S. Shin, T. S. Lee, et H. Kim, « An application of support vector machines in bankruptcy prediction model », Expert systems with applications, vol. 28, no 1, p. 127-135, 2005.

[31] Y.-C. Lee, « Application of support vector machines to corporate credit rating prediction », Expert Systems with Applications, vol. 33, no 1, p. 67-74, 2007.

[32] K. Kim et H. Ahn, « A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach », Computers & Operations Research, vol. 39, no 8, p. 1800-1811, 2012.

[33] H. Zhong, C. Miao, Z. Shen, et Y. Feng, « Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings », Neurocomputing, vol. 128, p. 285-295, 2014.

[34] X. Hu, J. Hu, L. Chen, et Y. Li, « Credit Risk Assessment Model for Small, Medium and Micro Enterprises Based on RS-PSO-SVM Integration », in 2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), avr. 2021, p. 342-345.

[35] S. Shetty et T. N. Vincent, « Corporate Default Prediction Model: Evidence from the Indian Industrial Sector », Vision, p. 09722629211036207, 2021.

[36] M. Moscatelli, F. Parlapiano, S. Narizzano, et G. Viggiano, « Corporate default forecasting with machine learning », Expert Systems with Applications, vol. 161, p. 113567, 2020.

[37] F. Z. Azayite et S. Achchab, « Hybrid discriminant neural networks for bankruptcy prediction and risk scoring », Procedia Computer Science, vol. 83, p. 670-674, 2016.

[38] H. Kim, H. Cho, et D. Ryu, « Corporate default predictions using machine learning: Literature review », Sustainability, vol. 12, no 16, p. 6325, 2020.

[39] M. Modina, F. Pietrovito, C. Gallucci, et V. Formisano, « Predicting SMEs' default risk: Evidence from bank-firm relationship data », The Quarterly Review of Economics and Finance, vol. 89, p. 254-268, 2023.

[40] M. Z. Abedin, C. Guotai, P. Hajek, et T. Zhang, « Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk », Complex Intell. Syst., vol. 9, no 4, p. 3559-3579, août 2023.

[41] A. Gicić et A. Subasi, « Credit scoring for a microcredit data set using the synthetic minority oversampling technique and ensemble classifiers », Expert Systems, vol. 36, no 2, p. e12363, avr. 2019, doi: 10.1111/exsy.12363.

[42] S. Tangirala, « Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm », International Journal of Advanced Computer Science and Applications, vol. 11, no 2, p. 612-619, 2020.

[43] Z. Zhao, T. Cui, S. Ding, J. Li, et A. G. Bellotti, « Resampling Techniques Study on Class Imbalance Problem in Credit Risk Prediction », Mathematics, vol. 12, no 5, p. 701, 2024.

[44] F. SASSITE, M. ADDOU, et F. BARRAMOU, « A Machine Learning and Multi-Agent Model to Automate Big Data Analytics in Smart Cities », International Journal of Advanced Computer Science and Applications, vol. 13, no 7, 2022.