Loan default prediction

Youssef Ashraf Kandil
Computer Engineering
The American University in Cairo
youssefkandil@aucegypt.edu

Abdullah Mohamed Kassem
Computer Engineering
The American University in Cairo
Abdullahkassem@aucegypt.edu

Introduction:

In the financial domain, the banking lending systems contribute to significant portions of profit. However, the loan application and approval processes have always been bottlenecks for a system; as they require numerous procedures, studies, and analyses. Since the loan profit depends entirely on the probability of return, human biases and errors are never acceptable. Thus, machine learning approaches have been studied and applied recently, to reach the finest models to minimize the loan defaults probability, as well as automate the whole lending process.

Problem specification:

Loan defaults probability prediction is a seriously difficult process since it demands analyzing a multitude of data about the loan recipient; considering a various range of not necessarily correlated variables. A loan application usually takes a few weeks to be approved or denied. Moreover, banks also struggle to build standards for loan acceptance, since it entails processing huge datasets that have been acquired throughout the years. Consequently, an optimal approach would be using a machine learning model to facilitate the data analysis.

Literature Review:

Literature review one:

Forecasting Loan Default in Europe with Machine Learning [1]

Brief description:

This research project studies and compares the driving factors for mortgage default in seven different European countries, by analyzing the data acquired in the Great Recession period. This project used several machine learning techniques to forecast load defaults, specifically for mortgages. The 12 million residential mortgages dataset utilized in this study was provided by the European Datawarehouse (ED), which is an entity that belongs to the European Central Bank (ECB). The dataset was subjected to Data cleansing to fix and eliminate errors before running the analysis. Here is a table demonstrating the type of data and features used in that project.

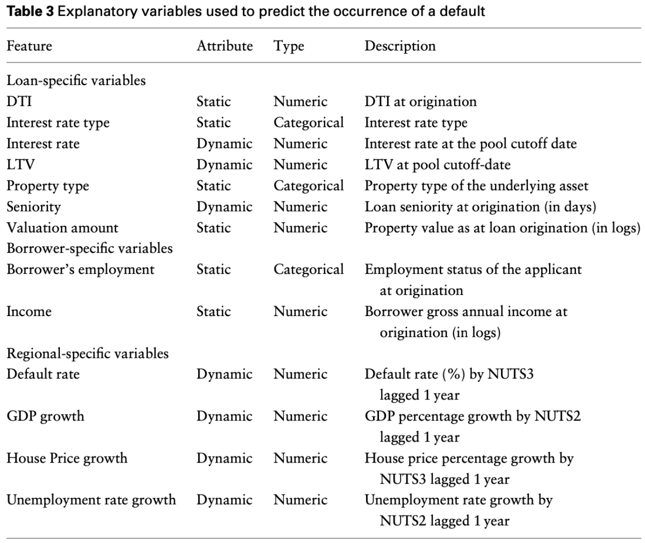


Table 3 Explanatory variables used to predict the occurrence of a default

Feature	Attribute	Type	Description
Loan-specific variables			
DTI	Static	Numeric	DTI at origination
Interest rate type	Static	Categorical	Interest rate type
Interest rate	Dynamic	Numeric	Interest rate at the pool cutoff date
LTV	Dynamic	Numeric	LTV at pool cutoff-date
Property type	Static	Categorical	Property type of the underlying asset
Seniority	Dynamic	Numeric	Loan seniority at origination (in days)
Valuation amount	Static	Numeric	Property value as at loan origination (in logs)
Borrower-specific variables			
Borrower's employment	Static	Categorical	Employment status of the applicant at origination
Income	Static	Numeric	Borrower gross annual income at origination (in logs)
Regional-specific variables			
Default rate	Dynamic	Numeric	Default rate (%) by NUTS3 lagged 1 year
GDP growth	Dynamic	Numeric	GDP percentage growth by NUTS2 lagged 1 year
House Price growth	Dynamic	Numeric	House price percentage growth by NUTS3 lagged 1 year
Unemployment rate growth	Dynamic	Numeric	Unemployment rate growth by NUTS2 lagged 1 year

Fig 1 [1]

Model used:

The following three machine learning models were used in this research project:

- Penalized logistic regression.
- Gradient tree boosting.
- Extreme GB.

Results and model comparison:

The dataset was divided into training, validation, and test sets, in respective percentages of 60%, 20%, and 20%. Since the data provided by (ED) is significantly unbalanced and shows a rare probability of loan defaults, the three sets were further manipulated to account for that unbalance. The sets were balanced by either under-sampling the most frequent classes or oversampling the less frequent ones. The results were analyzed using AUC and H-measure metrics for each of the three models with respect to each of the seven countries, and thereby, the boosting models showed more promising results than the logistic regression for all countries. The boosting models show very similar graphs, with a slight advantage to XGBoost in terms of H-measure.

Here is a graph showing the results.

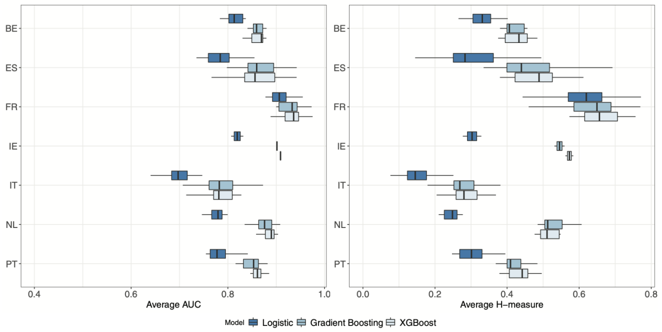


Fig 2 [1]

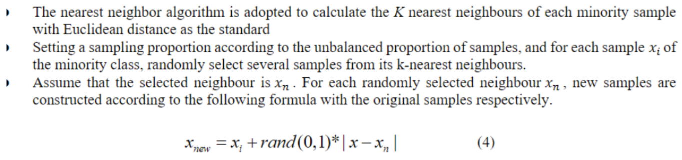
Loan default prediction dominant deriving factors:

The research did further analysis to the results identifying the dataset features which mostly contribute to the loan default prediction. Based on the analysis demonstrated in the paper, the LTV (loan-to-value) and interest rate data were the dominant features.

Literature review two:

A study on predicting loan default based on the random forest algorithm: [2]

This paper aimed to find the best performing machine learning algorithm for predicting loan defaults. It is important to note that our project goal is to create a machine learning algorithm that predicts bank loan defaults, while this paper deals with peer-to-peer loans. However, the training data is probably different as banks usually deal with higher loans than peer to peer, but still, this paper provides us with some significant information. Now we will learn how the problem was approached in this paper. First, the data set used was collected from a Lending club in 2019. It contained more than 115,000 loan data of users with 102 features; those numbers are within an acceptable range that should produce reliable results. As with most datasets, data cleaning was needed. Missing data was filled using mode interpolation, and the redundant and unnecessary features were removed. For feature engineering, new features were created using data from old ones, and feature abstraction was performed by encoding features using methods such as one-hot encoding. The last step of feature engineering was feature selection; the recursive Feature Elimination method was used to select 30 features from the 102 features with the strongest correlation with the target variable. Then the Pearson correlation graph was plotted, and some features were eliminated until the number of features reached 15. Finally, the Random Forest algorithm was used to determine and rank the most significant features, which reduces the learning difficulty giving us a more optimized model. A problem arises in the data, which is that 98.47% paid their loans, while only 1.53% were loan defaulters; this large difference could cause a problem in the model learning. So an oversampling method was adopted. The method is called SMOTE (Synthetic Minority Oversampling Technique).
The SMOTE is best described as:



- The nearest neighbor algorithm is adopted to calculate the K nearest neighbours of each minority sample with Euclidean distance as the standard
- Setting a sampling proportion according to the unbalanced proportion of samples, and for each sample x_i of the minority class, randomly select several samples from its k-nearest neighbours.
- Assume that the selected neighbour is x_n . For each randomly selected neighbour x_n , new samples are constructed according to the following formula with the original samples respectively.

$$x_{new} = x_i + rand(0,1) * | x - x_n | \qquad (4)$$

fig 3 [2]

After the model was trained using the random tree algorithm, the model was evaluated based on Accuracy, F1-score, Recall and ROC (receiver operating characteristic curve). Then these results were compared with 3 other machine learning algorithms, decision tree, SVM and logistic Regression.

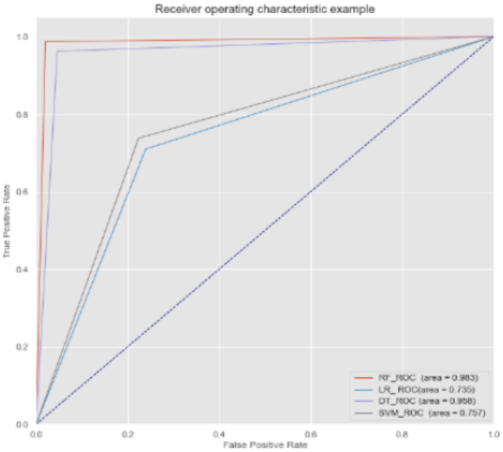


fig 4 [2]

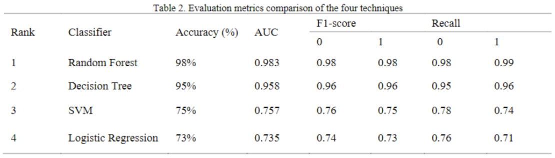


Table 2. Evaluation metrics comparison of the four techniques

Rank	Classifier	Accuracy (%)	AUC	F1-score		Recall	
				0	1	0	1
1	Random Forest	98%	0.983	0.98	0.98	0.98	0.99
2	Decision Tree	95%	0.958	0.96	0.96	0.95	0.96
3	SVM	75%	0.757	0.76	0.75	0.78	0.74
4	Logistic Regression	73%	0.735	0.74	0.73	0.76	0.71

fig 5 [2]

As it is visible from the ROC and the table, the Random Forest seemed to experience the best results and performance. The take from this paper is that the Random Tree algorithm could produce impressive results and is worth considering.

Proposed Procedure:

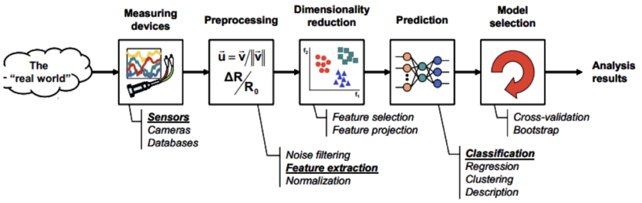


fig 6

By considering the approaches and better practices demonstrated in the reviewed literature, we chose datasets that are composed of raw data that have not gone through any kind of processing, in order to do the processing and cleansing from scratch. One of the three Datasets presented below in the next section, will be processed with respect to the diagram above. Through this project, we will be experimenting with some of the learned machine learning techniques to build an optimal loan default predictor, and formulate comparisons assessing the performance of the predictor. Logistic regression will be applied to assess the data and the performance of the other models since the two cited studies use it for benchmarking.

Datasets:

1. Dataset Name: Loan Defaulter:

https://www.kaggle.com/gauravduttakiit/loan-defaulter?select=application_data.csv

We found this dataset on Kaggle. It belongs to a case study where they apply Exploratory data analysis (EDA) on a set of data with 120 features and over 308k data samples. According to the user who submitted this data it is real; however, it has undergone data manipulation. The labels are well defined and explained. This dataset has a suitable number of samples, enough to divide for testing and training, and many features. Also this dataset has been used by a number of notebooks on Kaggle, that means that some people think it is usable and their notebooks could act as a reference to us. Since this data belongs to a case study where they use EDA, it leads us to believe that EDA would be a good approach to deal with this data. The features used are a lot and logically would seem to make a difference in whether the person will be able to pay the loan or not.

2. Dataset Name: Loan Prediction Based on Customer Behavior:

https://www.kaggle.com/subhamjain/loan-prediction-based-on-customer-behavior?select=Training+Data.csv

This Dataset belongs to a Hackathon organized by Univ.AI. The inspiration behind this dataset is that an organization wants to predict whether the user who is applying for a loan is risky or not. This dataset provides us with information about the user's history such as their income, ownership of car etc and for each sample in the dataset there is a column that contains 0 if they paid their loan and 1 if they defaulted the loan. The dataset has 11 features and 252k samples for training and 28k for testing.. This dataset has enough samples for testing and training. It does have less features than the other dataset, which is a minor drawback. Although a lot of people used this dataset in their Notebooks, some left comments saying that this data is "randomly labeled".

3. Dataset Name: Loan Default Dataset

https://www.kaggle.com/yasserh/loan-default-dataset
This dataset, as the others, contains many features about the loan and the person asking for the loan. It has around 32 features and 149k samples. Which is enough both for testing and training. According to the author of the dataset it contains a lot of multicollinearity & empty values. Which makes it difficult to handle and work with. This caught our attention because good and easy to handle data will not always be available, so working with this data would be a good challenge. A drawback of the dataset is that there is no clear description of the features and some are a bit hard to understand.

Bibliography:

1. Luca Barbaglia, Sebastiano Manzan, and Elisa Tosetti. 2021. Forecasting loan default in Europe with machine learning. (July 2021). Retrieved February 13, 2022 from https://academic.oup.com/jfec/advance-article/doi/10.1093/jjfinec/nbab010/6313398#267157111
2. Lin Zhu, Dafeng Qiu, Daji Ergu, Cai Ying, Kuiyi Liu. 2007. A study on predicting loan default based on the random forest algorithm, Procedia Computer Science, Volume 162, Pages 503-513, ISSN 1877-0509. DOI: https://doi.org/10.1016/j.procs.2019.12.017.
3. https://www.kaggle.com/gauravduttakiit/loan-defaulter?select=application_data.csv
4. https://www.kaggle.com/subhamjain/loan-prediction-based-on-customer-behavior?select=Training+Data.csv
5. https://www.kaggle.com/yasserh/loan-default-dataset