Machine Learning for Loan prediction and analysis for publication in *Journal of Physics: Conference Series*

Lic. Arnoldo Oliva

E-mail: hugo.olivac@uanl.edu.mx

Abstract. Banking institutions depend greatly on the lending activity, as this allows them to charge the customer an additional rate of the money they lend. At the same time, this fact represents a risk of loss as the client is not guaranteed that he will pay all the charged quantity. Even so it would be more difficult for a company to not pay a loan, this is also the case for those entities. Efforts have been made in this industry to develop Machine Learning models that grant more certainty of evaluating a more accurate possibility of a certain customer of not paying the charged loan, thus minimizing the risk for the banking entity of losing. In this paper ML models are deployed to help to the acheivement of this goal, resulting in some interesting findings.

1. Introduction

This paper is an attempt to discover important factors regarding loan default and also attempts to develop a model that predicts with high accuracy the probability of falling in default category. To achieve this, a dataframe was worked on, originally from the SBA (USA Small Business Administration Office).

The goal of the SBA is to encourage small businesses to access to the credit market, and one tool it has to comply with this task is a SBA loan guarantee programs. Another effort to comply with this is the tracking of the loans and their outcomes, either fully paid or default, resulting on the SBA Administration dataset. This is important to track, not only for banking institutions, because as economic theory suggest, small businesess are a primary source of job creation, thence, *"fostering small business formation and growth has social benefits by creating job opportunities and reducing unemployment".* Li et al. [2018].

2. Literature Review

This section discusses in brief about some of the work that has already been done on creating ML and DL models using various algorithms to improve the loan prediction process and help the banking authorities and financial firms select an eligible candidate with very low credit risk.

According to the paper *"Should This Loan be Approved or Denied?": A Large Dataset with Class Assignment Guidelines* from Min Li, Amy Mickel and Stanley Taylor (2018), in which the authors studied the same dataset as this paper, the following procedure should be taken into account when predicting probabilities of a loan falling in default:

A logistic regression should be considered to determine that probability as it provides easy understanding outcomes, it gives great insights, and also can work with binary variables, opposite case of the linear regression model.

The authors applied interacting variables to capture special phenomenons, and also guided by economic theory most of all to decide which variables to use.

An example of this is that they built a Recession dummy variable which indicates whether or not the loan falls in the economic crisis period of time (2008-2010); a *term* variable which means that if the month term of the loan is greater than 240 months, then the loan is covered by real state, among other examples.

The authors limited their work to only one state and only one NAICS category alleging that, if a person wanted to work with multiply categories, estimations errors could be possible.

When evaluating the results, a mean absolute error of 32 percent was obtained according to the results of the confusion matrix.

In the present paper, as a goal to isolate economic crisis variables, to comply with the assumption of normal economic cycle (growth) and also excluding state factor (the data contains records for all US states).

To amplify the scope of our study regarding more complex analysis regarding ML models like Random Forests models, the paper *Loan default prediction using decision trees and random forest: A comparative study* from Madaan et al. [2021] was also reviewed. In this investigation, authors focused on more personal data, contrasting with the scope of this paper which is business loans. Authors acknowledge the importance for the banking system of deploying models that help to give a score of the risk of falling into default of the customers, as this may lead to losses. Based on the work of another study, authors noted that, among the results of multiple ML models like logistic regression, support vector machine, decission trees, Random Forests models take the lead regarding accuracy. On the contrary, another paper that they reviewed stated that SVM models are superior in prediction than the RF models. Authors decided to work with Decission Tree and Random Forest models, and RF models mainly because of the immunity to overfitting, the accuracy, and efficiency on large datasets.

Mehul Madaan et al, before deploying the model, did some exploratory analysis and found interestings insights like that one of the most asked loans is the funding of small businesses, followed by home improvement. To develop the model, they splitted the data into a 70-30 distribution of training and testing, and set an estimator count of the RF of 100. At the end, they compared the results between the two models, resulting that the accuracy of the RF model with 80 percent outperformed the Decission Tree model with 73 percent of accuracy.

3. Methodology

3.1. Data processing

Data was treated in a way it can be compatible with ML models, and also involves filtering of non relevant components. Some examples of that were that for this case study, as stated before, the dates were selected to avoid the economic crisis period from 2007 to 2011. The loans that were selected complied with approvement and disbursement dates between the year 2000 and 2006. Also, geographic variables were dropped following the assumption of most basic economic variables, *zip*, *city*, *states*. Geographic effect was captured by a dummy variable *same state*, which indicates if the bank is on the same state as the business, and *urban rural* which indicates if a business is in a rural or a urban environment.

Timestamps of approvement and disbursement dates were obtained to check for the time difference between those two variables. Data categorization, data cleaning and data replacing were made to adjust the data. As the final step, the continuous variables were scaled and the categorical variables were converted to dummies.

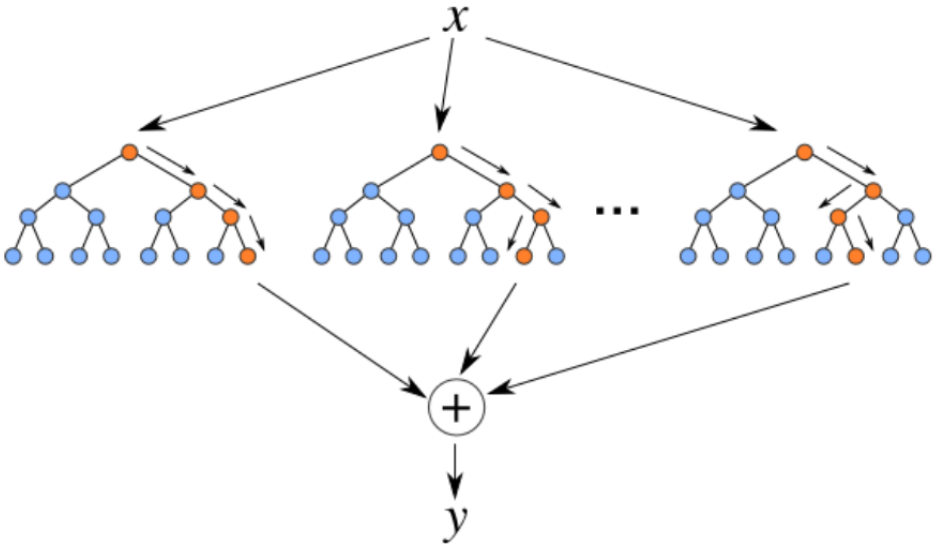
The result was a dataset of 313 k records compared with the original of 899 k records, with the following continuous variables scaled: month terms of the loan (Term), number of employees (NoEmp), amount of disbursement gross (DisbursementGross), gross amount outstanding (BalanceGross), charged off amount (ChgOffPrinGr), gross amount approved by bank (GrAppv) and gross amount approved by the SBA (SBAAppv), and the time stamp difference between Approval date and Disbursement date of the loan. Some of the categorical variables were treated as *dummies*, a dummy variable is dicotomical manner to display categorical variables, where it represents the absence or presence of a condition of a category of the variable. The dummy variables that are being worked are: NewExist, which indicates to us if the business is new or has been operating for some time, UrbanRural which indicates whether the business is located on an urban community or not; if the business has the Revolving Line credit option, SameState which indicates if the business is located on the same state as the bank they are being borrowed. Low Doc Loan program, if the business has created jobs, if the business has retained job positions, and dummy variables for the following NAICS categories: Construction, Health care and social assitance, Manufacturing, Retail, Professional scientific and technical services and Other.

3.2. Supervised learning methodology

As the problem faced on the scope of this study is to predict default loans and not only obtaining insights, a Supervised learning method was taken into account. With the data treated for the unsupervised section, a Random Forest classifier algorithm was tested on the data, as Madaan et al. [2021] suggested this method can result on decent accuracy scores, besides of the fact that allows to display variable importance alike results.

3.2.1. How does a Random Forest classifier algorithm works? Decision Tree algorithm concept works as follows: for a particular variable there is a splitting criteria node that separates the data in two branches, repeating this process until a node with no more possible branches is reached. Random Forest algorithm is based off on multiple decision trees, that each of those trees use a different sample of the data. In the end, the RF model merges and summarizes all the decision trees to classify the outcome, in this case, a binary response one.

According to R [2023] a RF model is an ensemble technique, meaning that it combines multiple models to achieve a goal. *Bagging* is the ensemble technique used on RF, in which this approach chooses a random sample from the entire data set, with replacement. Each model then is trained separately based on these selections.



A visual representation of the RF algorithm, courtesy of Harp

To define the mathematics of the branch splitting of the nodes in a Random Forest classifier algorithm, it is worth noting that the default option selected in *SkLearn Python* library that will be used for this work is the Gini impurity index, which can be represented as the following:

$$Gini = 1 - \sum_{i=1}^{C} p_i^2 \tag{1}$$

Where C is the number of classes and p_i is the probability of selecting an item from class i.

This formula uses the classes and probabilities to determine the Gini coefficient, determining which of the branches will be more likely to occur. In the formula, pi represents the relative frequency of the particular class in the data set sample and c is the number of all classes.

Some pros according to Schott are that the RF models can prevent overfitting the data as each decision treee is based off on a different, independent sample of data, and some cons are that these models mainly are very sensitive to outliers and holes in the data, and also that they can be slow to train in large data sets.

Coming back to the methodology of this paper, it is worth noting that the processing of the data selected resulted on an unbalanced set, with around an 82 percent of the cases with totally paid loans. Data was balanced with the following procedure: All failed loans were selected, and a random subset of the equivalent quantity of failed loans (55,942) was made on the successful loans, having the data balanced with a proportion of 50-50.

Thenceforth, data was splitted on training and testing sets with a ratio of 80-20. The idea of this is to have the algorithm training on a bigger dataset and then testing the results on another. A random forest classifier model was made with a number of 100 decission trees, resulting on an accuracy index of 99.57 percent. A cross validation technique of 10 iterations was made to ensure the model was not over fitting the results, and the mean accuracy resulted on 99.59 percent, with a standard deviation of 0.0005, meaning that the model accuracy was consistent besides of the random sampling evaluation technique.

Regarding the insights of this algorithm, we use a *Feature Importance (FI)* option from the RF built in method in Python. As the RF model is based on Decission Trees, which are composed by nodes, and within each node the selected feature divides the data into two separate subsets with similar characteristics within. Each feature is selected by a Gini impurity criteria, and for each feature we can collect the average of how that variable decreases this impurity. This would be the Feature Importance, which can serve us as a proxy of how much each variable determines the outcome of the model.

Sorted in terms of FI (the sum of all of them has to be the 100 percent), "ChgOffPrinGr" is the most important with 77.85 percent. This makes sense as this variable shows the charged off amount (amount that the lender has set as loss); the second most important is "Term", which is the loan term in months. The more time you have, in theory the easier is it to pay it. The third one is "SameState" with 1.3 percent, and the following would be "SBA Approved", with 1.1 percent. The rest of the variables contain very few percentages. A second iteration of this algorithm will be run without the ChgOffPrinGr as this variable would be very related to the "Paid" feature.

Comparing the results obtained from Madaan et al. [2021] as these authors worked on a Random Forest algorithm to predict whether an individual should be given a banking loan, they obtained an accuracy of 80 percent, contrasting with the first iteration of RF model in this paper of 99.57 percent. The first thing to note is that the differences in the approaches may lead to take different selection of data selection and data treatment, so it is not a great comparison. To contrast another approach with a more feasible analogy, Li et al. [2018] paper work was also contrasted, as the authors selected the same data set, for the same purpose of

predicting failure paying loans. Recalling that Li et al. [2018] used a Logistic regression model to obtain more insights, authors defined a 0.5 cutoff probability, obtaining an accuracy of levels around 68 percent. It is worth recalling that these authors took a very different approach as they decided to use variables encapsulating the economic crisis period, disregard using the NAICS categorization and isolated their work only for the state of California, USA.

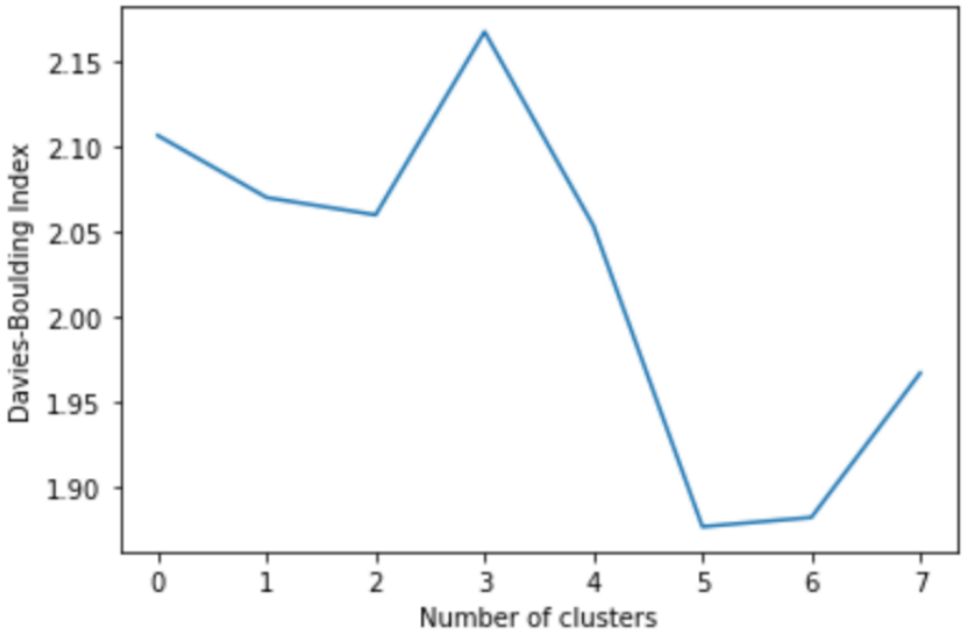
3.3. Extra: Unsupervised learning

A first iteration of unsupervised learning with the objective of identifying possible patterns of data that may affect the outcome of the loan outcome was made. Before talking about this, its worth noting that some filter methods were tried to select the most important variables that would be part of the unsupervised learning clustering technique (in this case, was Kmeans algorithm).

A variance threshold model was ran in order to obtain this objective. Mainly all of its inputs were continuous variables.

A Lasso regression was made as this regression allows to work with categorical and continuous data at the same time (in this iteration the dummy variables for the most important categories of the NAICS were took as input). Both methods resulted in very different outcomes due to their differences in the methodology, but also got some similar results: variables like ChgOffPrinGr, Term, timediff, SBAAppv were marked in both techniques as the most important ones; but DisbursementGross, GrAppv, RevLineCr,SCIAN category (without being dummy) and NoEmp were either in one model or another, but not in both.

With continuous data being scaled, and categorical data (mainly most important categories of NAICS) being treated as dummies, an iteration of the KMeans algorithm was ran to identify the best number of clusters that reduces the most the Davies-Bouldin index. For this case was the model with 7 clusters reducing this value up to 1.87. In the following image there is an example of the evolution of the Davies-Bouldin index.



References

Min Li, Amy Mickel, and Stanley Taylor. Should this loan be approved or denied?: A large dataset with class assignment guidelines. *Journal of Statistics Education*, 26(1):55–66, 2018. doi: 10.1080/10691898.2018.1434342.

Mehul Madaan, Aniket Kumar, Chirag Keshri, Rachna Jain, and Preeti Nagrath. Loan default prediction using decision trees and random forest: A comparative study. *IOP Conference Series: Materials Science and Engineering*, 1022(1):012042, jan 2021. doi: 10.1088/1757-899X/1022/1/012042.

Sruthi E R. Understand random forest algorithms with examples (updated 2023). https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/, 2023. Accessed: February 26, 2023.

Harp. Harp random forests. https://dsc-spidal.github.io/harp/docs/examples/rf/. Accessed: February 26, 2023.

Madison Schott. Random forest algorithm for machine learning. https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9 Accessed: February 26, 2023.