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.

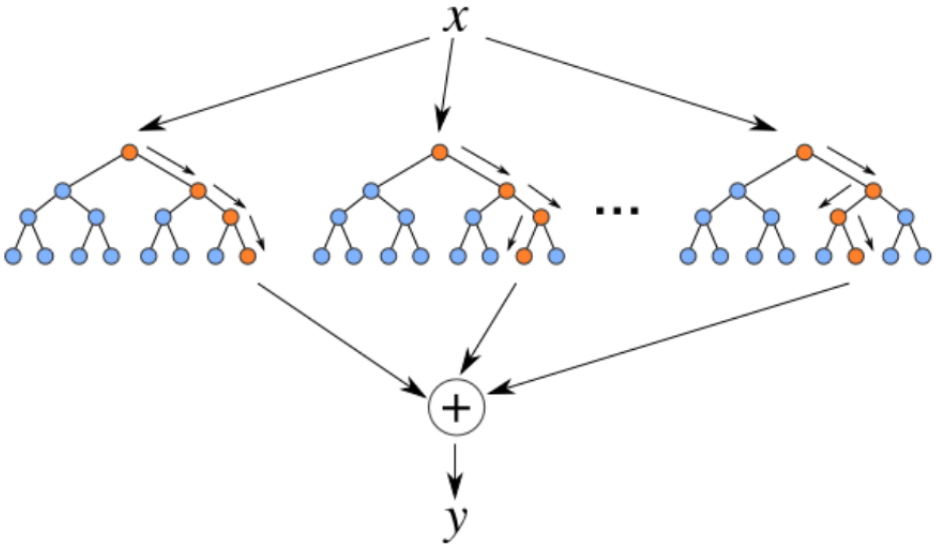
The result was a dataset of 313 k records compared with the original of 899 k records, with the following continuous variables scaled: month terms of the loan (Term), number of employees (NoEmp), amount of disbursement gross (DisbursementGross), gross amount outstanding (BalanceGross), charged off amount (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 records) 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. In Section 5, an experiment is conducted with non scaled data of "ChgOffPrinGr" as in that section is described that that variable is highly related in economic theory with the Paid outcome, henceforth on this section the model is ran without that variable, and also worth noting that is ran with scaled data.

To get insights with this algorithm, we use a *Feature Importance (FI)* option from the RF built in method in Python. As the RF model is based on Decision 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), "Term" is the most important with 58.40 percent. This makes sense as this variable indicates the time in months that the borrower business has available to pay; the more time you have, in theory the easier is it to pay it. Following variables have way less importance than the "Term" variable. The second most important variable is the Time difference between the Disbursement timestamp (date when the money was delivered for lending) and the Approval timestamp (date when the loan was approved as such), with almost 8 percent. Next most important one was "SBAAppv", which indicates the SBA's guaranteed amount of approved loan with 6.5 percent. And very close, DisbursementGross (amount disbursed) is following with 6.3 percent. Gross amount of loan approved by bank (GrAppv) is the last one in the top 5 most important variables with 5.2 percent, and Number of employees has a 4.6 percent of importance. The sum of those 6 variables represent the 88.3 percent of the model. The rest of the variables contain very few percentages.

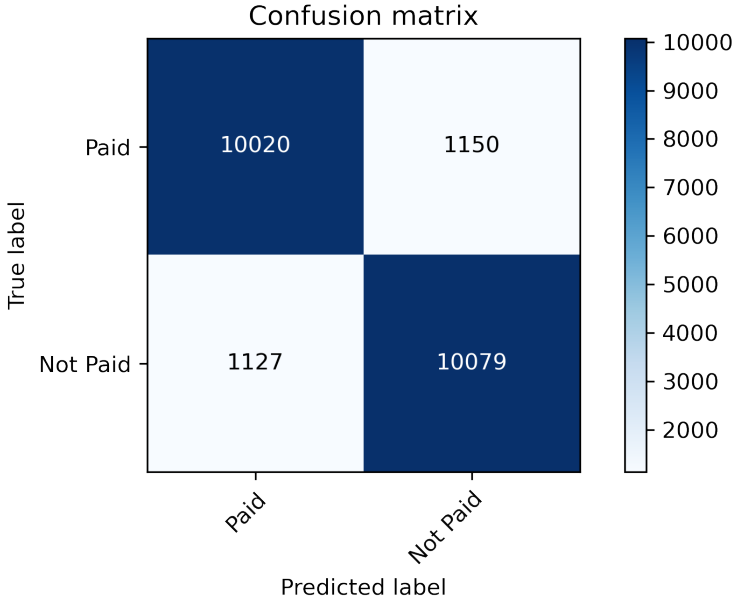
4. Results

For this section the methodology of the papers reviewed are contrasted in order to decide which metric to use to evaluate the efectiveness of the model. To contrast another approach with a

more feasible analogy, Li et al. [2018] paper work was contrasted, as the authors selected the same data set, for the same purpose of predicting failure paying loans. It is worth noting 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. Recalling that Li et al. [2018] used a Logistic regression model to obtain more insights, authors defined a 0.5 cutoff probability based on the fact that misclassification rate was the optimal at that level. Authors made a confussion matrix, and decided to evaluate the results of their algorithm with a "missclasification rate" in which they sum the values of the False negatives and False positives, which practically would be the reverse of the accuracy metric. They got an accuracy metric of 67.84 percent.

As Madaan et al. [2021] worked on both a Decission Tree algorithm and a Random Forest algorithm (for practical purposes, in this ocasion we will be only taking into account the RF model) to predict whether an individual should be given a banking loan, their metric should be taken into account also. Authors also worked with accuracy metric, which is the proportion of values that were correctly labeled compared with the actual value (True positives TP + True negatives TN), divided by all type of values (including TP, TN, and the missclasified values False Positives, or Error Type I, and False Negatives, or Error Type II). The authors obtained an accuracy metric of 80 percent, contrasting with the Random Forest model results in this paper of 89.82 percent. The first thing to note is that the authors selected another dataset, and also differences in the approaches may lead to take different selection of data selection and data treatment, so it is not a great comparison.

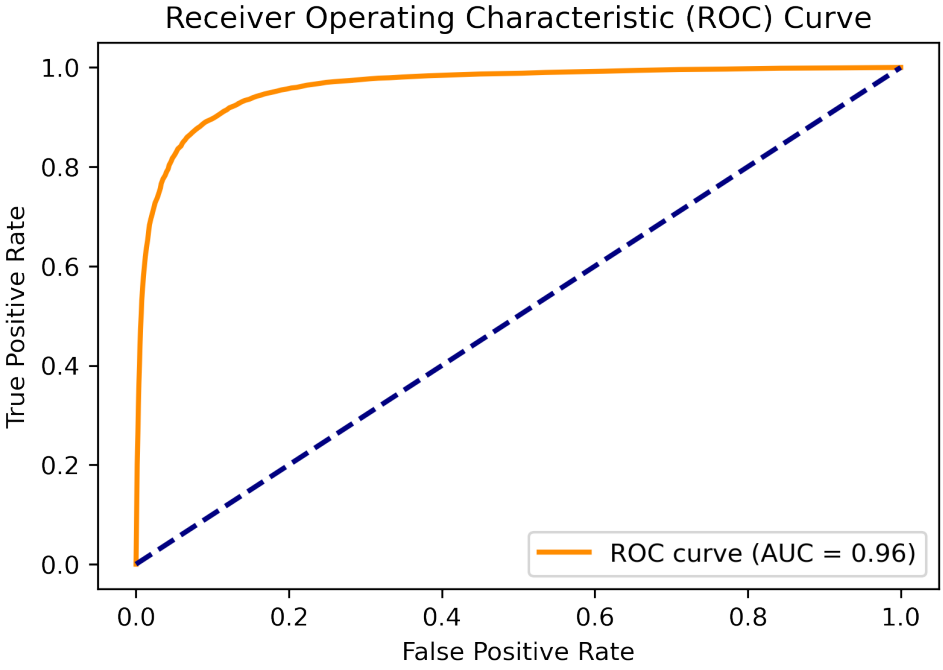
As reviewed papers used accuracy metrics, it will be interesting to use also this metric, but in addition, another metric can be used. In the first iteration of this work, an accuracy metric of 89.82 was obtained. Please find below the confusion matrix for this iteration (worth remembering that it was constructed over the test sample containing the 20 percent of the working sample, or 22,376 records):



A mean of an accuracy metric of 88.95 percent was obtained in a cross validation technique of 10 folds. Cross-validation is a technique used in machine learning and statistical modeling to evaluate the performance and generalization ability of a predictive model. The main purpose of cross-validation is to estimate how well a model will perform on new and unseen data. This technique involves partitioning the available data into multiple subsets or "folds", typically k

folds. The model is then trained on k-1 folds and tested on the remaining fold. This process is repeated k times, with each fold being used as the test set once. The results of each iteration are then combined and summarized to produce a final estimate of model performance. Thus, this result indicates that this model is capable of doing effective predictions even with new sample data, and it is also a good indicator of the absence of overfitting.

There is also another indicator of the effectiveness of the model, which is called *ROC AUC Curve*. The ROC (Receiver Operating Characteristic) curve is a graphical representation of the performance of a binary classifier system as the discrimination threshold is varied. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The TPR is the proportion of actual positive cases that are correctly identified as positive by the classifier, while the FPR is the proportion of actual negative cases that are incorrectly identified as positive. The TPR is also known as sensitivity or recall, while the FPR is equal to 1 minus the specificity. The AUC (Area Under the Curve) of the ROC curve is a metric that summarizes the performance of the binary classifier over all possible thresholds. It represents the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. An AUC of 0.5 means that the classifier is no better than random, while an AUC of 1.0 means that the classifier is perfect. In general, a higher AUC indicates a better classifier performance. The ROC curve and AUC are useful for evaluating and comparing the performance of different binary classification models. Please find below a ROC AUC curve graph for this model:



An AUC ROC (Area Under the Receiver Operating Characteristic Curve) of 0.96 is a very good result, as it indicates that the classifier has excellent discriminatory power and can effectively distinguish between the positive and negative classes.

In general, an AUC ROC score of 0.5 indicates random guessing, while a score of 1.0 indicates perfect classification. Therefore, an AUC ROC score of 0.96 indicates that the classifier is achieving very high accuracy, sensitivity and specificity.

In the next subsection an experiment is going to be designed to contrast how that particular variable of "Charged off amount" impacts the results of the dependent variable.

5. Experiment designing

An experiment represents a change in the conditions of the study problem. The problem involves independent and dependent variables. For the design of experiments, levels and treatments of these variables are considered. The levels are the values that the independent variables take, while the treatments are the unique values that are assigned to each combination of levels in the problem.

The experiment will involve the Charged-off amount (ChgOffPrinGr) variable, as this variable is highly related to the dependent variable in the economic theory: this refers to a quantity of the loan that the creditor considers as uncollectable.

For the experiment, the levels are going to be constructed in the independent variable of Charged-off amount, and the treatments over the dependent variable. It is worth noting that for this experiment, the non scaled dataset will be selected as the min max scaling for the Charged-off amount resulted in troubles for the python algorithm as there were too many values extremely close to zero. This may lead to a different result in the accuracy index.

Having the levels constructed over based on the median (50 percent quantile) and 75 percent quantile of the ChgOffPrinGr variable, in where all values below the median were labeled as "Low charged off amount", between the median and 75 percent quantile as "Medium Charged off amount" and above the 75 percent as "High Charged off amount", a proportion test of Paid cases was conducted.

To perform a proportion comparison test to verify that the independent variable affects the dependent variable (by comparing its multiple constructed levels), the "proportionschisquare()" function from scipy.stats can be used to conduct a chi-square test to compare proportions across k samples. This function uses the chi-square test and the chi-square distribution to perform the test. The chi-square test for proportions does not require the distribution of the samples to be normal, but some conditions need to be met. Specifically, it is assumed that the samples are independent (which they are because one sample does not affect the other) and that the proportion of successful cases in each sample is sufficiently large (greater than 40). In the chi-square test for proportions, the null hypothesis is that there are no significant differences between the proportions of the k samples. That is, the proportion of successful cases in each sample is equal. If the p-value is less than the chosen level of significance, it can be concluded that there are significant differences between the proportions of the k samples.

To do the previous description of the test, a contingency table of Paid cases was constructed depending on the levels of Charged off amount, with the following results:

	Paid: 1 (success)	Paid: 0 (failure)
Low charged off amount	55515	427
Medium charged off amount	380	27591
High charged off amount	47	27924

Table 1. Table of charged off amounts and payment status

Continuing with our test, The null hypothesis can be written as:

$$H_0 : p_1 = p_2 = \cdots = p_k$$

where p_i represents the proportion of successes in the ith sample.

Let $\alpha = 0.05$ be the chosen significance level.

To test this hypothesis, we can use the chi-square test for proportions. The test statistic is calculated as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed number of successes in the ith sample, and E_i is the expected number of successes in the ith sample under the null hypothesis. The expected number of successes can be calculated as:

$$E_i = n_i \cdot \hat{p}$$

where n_i is the sample size of the ith sample, and $\hat{p}$ is the pooled proportion of successes across all samples, calculated as:

$$\hat{p} = \frac{\sum_{i=1}^{k} O_i}{\sum_{i=1}^{k} n_i}$$

Under the null hypothesis, the test statistic follows a chi-square distribution with $k-1$ degrees of freedom. We can calculate the p-value as the probability of observing a test statistic as extreme as the one calculated under the null hypothesis. If the p-value is less than the significance level α, we reject the null hypothesis and conclude that there are significant differences between the proportions of the k samples. Otherwise, we fail to reject the null hypothesis.

Continuing with our test, after deploying the code in python, these are the following results: with a Degrees of Freedom of 2 (3 samples minus 1), a p value of 0.0 was obtained, and contrasting with our *alpha* of 0. (of 95 percent of conficende), we note that the Null Hypothesis of no significant differences between the proportions of the k samples is rejected as the p value was smaller than our alpha value. Hence, there are significant differences of the proportions of Paid cases between the different levels of Charged off amount; Charged off amount impacts directly on Paid/Non Paid results.

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.