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

Lic. Hugo Arnoldo Oliva Castillo

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 Machine Learning models are deployed to help to the achievement of this goal, resulting in some interesting findings.

1. Introduction

This paper seeks to discover important factors regarding loan default but mostly seeks to develop a model that predicts with high accuracy both categories of paid and unpaid cases of loans. 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, in which SBA acts like an insurance provider, with the objective to reduce the risk for the bank by taking on some of the risk of the credit through guaranteeing a portion of the loan. In the particular case that the loan is not paid (goes into default), SBA covers the amount that they initially guaranteed. Li et al. [2018].

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. There are plenty of reasons why it is important to track credit scores, both for banking institutions, clients either individuals or businesses, and the economy as a whole. Having a good credit score is crucial for individuals who want to secure loans and other financial services from banks. A credit score is a numerical representation of an individual's creditworthiness, which is based on their borrowing and repayment history. Banks and other lenders use this score to determine whether or not they should lend money to someone and at what interest rate.

For individuals/businesses, maintaining a good credit score is important because it can help them obtain loans for various purposes, such as buying a home or expand the capabilities of

the business. It can also make it easier to get credit cards, qualify for lower interest rates, and negotiate better terms for loans.

For banks, focusing on the creditworthiness of their clients is also crucial. By ensuring that they lend money to individuals with good credit scores, banks can reduce their overall risk and increase the likelihood of receiving timely loan repayments. This can help them maintain a healthy financial position and avoid losses due to defaults or delinquencies.

In addition, by providing loans to individuals/businesses with good credit scores, banks can help them achieve their financial goals and improve their overall financial well-being. This can lead to a stronger economy, as individuals with access to credit are more likely to make investments and take risks that can drive economic growth. 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].

Overall, it is in the best interest of both businesses and banks to focus on building and tracking credit scores. By doing so, businesses can access the financial services they need to achieve their goals, while banks can mitigate their risk and support economic growth.

This descriptions over the ideal credit system sounds great in theory, but in the real world is not necessarily like this. Although in this database, it was filtered to check the loans over a period of time with no economic crisis happening at the same time to isolate a normal economic cycle (assuming sustainable economic growth), there were still the 17.86 percent of the records with cases of default loans. This may sound few compared with the other percent of succesfully paid loans, but the effects of large unpaid cases may have devastating effects in the financial institutions and the economy.

Unpaid loans can lead to significant losses in revenue and profitability. This is because banks typically rely on the interest earned from loans to generate income. When loans are not repaid, banks may have to write them off as losses, which can have a negative impact on their balance sheets and financial stability. In extreme cases, unpaid loans can even lead to the failure of the bank. On a broader level, unpaid loans can have negative effects on the economy as a whole. When individuals default on loans, it can reduce the amount of money available for lending, as banks become more risk averse. This can make it more difficult for businesses and individuals to obtain credit, which can lead to slower economic growth.

That are some of the reasons why it is important to develop machine learning (ML) models that helps the banking institutions to detect the chances a loan will be defaulted, and the reasons of this event. This may gave financial institutions tools to prevent the losses described earlier. The motivations to develop effective ML models can help the banking institutions and the economy as a whole, are some of the reasons of the developing of the current paper.

2. Literature Review

This section discusses in brief about some of the work that has already been done on working ML algorithms to improve the loan prediction outcome and help the banking authorities and financial firms select an eligible candidate with adequate 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.

The authors considered that a logistic regression should be taken into account to determine the probability of paying/not paying loans, 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 they were

guided by economic theory 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). Authors wanted to explore the variations of the loans of before and after the Great Recession which took place on those years. The authors believed that loans with disbursement date after 2010 would provide more weight to charged off loans compared with fully paid loans. Although, authors recognized that this procedure would create the possibility of selection bias in the end of time period. Additionally, in their interacting variables list, they created a *term* variable which means that if the month term of the loan is greater than 240 months, and if the loan is covered by real state guarantees, among other examples.

In this dataset, it is captured the economic activity of the business with the NAICS. NAICS stands for "North American Industry Classification System". It is a system used in Canada, Mexico, and the United States to classify businesses into different industry sectors based on the type of economic activity they engage in. The authors limited their work to only one state (California, U.S.) and only one NAICS category (real estate rental and leasing) alleging that, if a person wanted to work with multiply categories, there was a probability for estimations errors to happen. Authors selected randomly half of the data for training data set. Their dependent variable was set as 1: Having default loan, and 0: having fully paid loan, to predict possibility of unpaid loan. When evaluating the results, a mean absolute error of 32 percent was obtained according to the results of the confusion matrix. Authors recognized that a higher accuracy can be reached with more sophisticated ML algorithms and techniques. Their results obtained from the model indicates that, for that sample selected, RealEstate dummy (having 1 if its backed by real estate) was significant with negative coefficient, meaning that if its backed by real estate, it is less probable to have a unpaid loan, Portion guaranteed by SBA, also with negative coefficent and statistically significant, and Recession dumy variable, statistically significant but with possitive coefficient, meaning that loans happening in Recession times have bigger chances to fall in default. The dummy variable of being a New Business and the Disbursement Gross amount resulted on being non statistically relevant to the model (their p value exceded by far the 0.05 alpha value of significance).

In the present paper, as a goal to isolate economic crisis variables, we select the data from the year 2000 to 2006 to comply with the assumption of normal economic cycle (in those years economic growth was present, Gross Domestic Product overall was growing in the world), and also, the present work will exclude state factor (the data contains records for all US states), as is it not an objective to explore the coyuntural socioeconomic differences between the 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 individual 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 other studies, authors noted that, among the results of multiple ML models like logistic regression, support vector machine (SVM), decission trees (DT), Random Forests (RF) models in general 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.

Madaan et al. [2021], 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 Random Forest of 100 trees.

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. Authors discovered some interesting conclusions such as the combination of a non-homeowner applying for a loan for a small business or wedding may lead to the borrower defaulting on the loan, which could have a negative impact. Madaan et al. [2021].

Although of the differences of the data sets, the current one being focused on business lending and the authors focused on the purposes of individual lendings, this paper was analyzed to review some of the strengths of using a Random Forest model, compared to other algorithms. The authors established that the RF technique can outperform some other algorithms. In the next section this technique is going to be performed over the SBA business lending data set to check for the accuracy levels of this technique and also to gain some interesting insights with another approach compared to Li et al. [2018].

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 *"zip"*, *"city"*, *"states"* were dropped to isolate the analysis excluding geographic coyuntural data, to work following the assumption of basic economic data. Even though, a geographic effect was captured by a dummy variable *"same state"*, which indicates if the bank is on the same state as the business, to explore if the location of the business in regard to the bank can have an impact over the loan payment process; 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, and they were captured in a variable called *"timediff"*. Data categorization, data cleaning and data filtering were made to adjust the data. As the final step, the continuous variables were scaled to avoid the possibility of one variable weighting more than the others; and the categorical variables were converted to dummies to explore the impact of the presence of each variable over the dependent variable *"paid loans"*.

The result was a dataset of 313 k records compared with the original of 899 k records (as we select a subset of the data comprising dates from the year 2000 to 2006), 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 of the loan approved by bank (*"GrAppv"*) and SBA's guaranteed amount of approved loan (*"SBAAppv"*), and the time stamp difference between Approval date and Disbursement date of the loan. Those were the non categorical variables. 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 (*"CreateJob"*), if the business has retained job positions (*"RetainedJob"*), and dummy variables for the following NAICS categories: *"Construction"*, *"Health care and social assitance"*, *"Manufacturing"*, *"Retail"*, *"Professional scientific and technical services"* and *"Other"*. This particular NAICS categories where selected as they where the classes with more records compared with the rest.

3.2. Supervised learning methodology

As the problem faced on the scope of this study is to predict default loans, a Supervised learning method was taken into account. Supervised learning technique refers to a ML problem where the outcome is already known, so we train the model with the known answers. More particularly, 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.

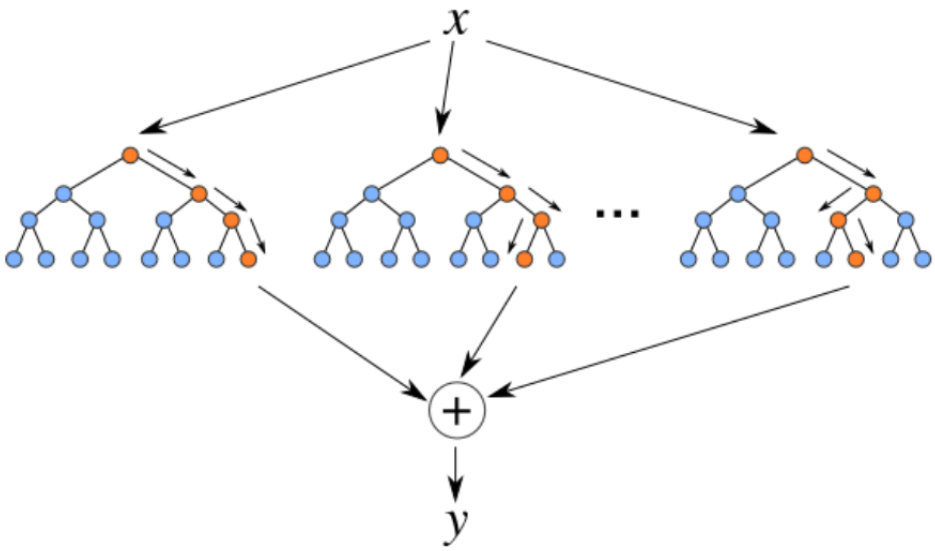

Figure 1. RF algorithm

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:

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

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.

3.2.2. Implementation of Random Forest classifier 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 fully paid loans (55,942 records) was made, 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 the model 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 7.88 percent. Next most important one was *"SBAAppv"*, which indicates the SBA's guaranteed amount of approved loan with 6.53 percent. And very close, *"DisbursementGross"* (amount disbursed) is following with 6.39 percent. Gross amount of loan approved by bank (*"GrAppv"*) is the last one in the top 5 most important variables with 5.23 percent, and Number of employees has a 4.69 percent of importance. The sum of those 6 variables represent the 89.12 percent of FI of the model. The rest of the variables contain very few percentages.

3.3. Support Vector Machines

As Madaan et al. [2021] recognized in their Literature Review that Support Vector Machines (SVM) can deploy good accuracy indicators, in this paper this technique is also added. According to Huang et al. [2018], SVM is a method that work as a classifier by creating a decission boundary between two or more classes enabling the prediction of labels from one or more feature vectors. This decission boundary, also known as "hyperplane", is positioned in a manner that maximizes the distance from the nearest data points of each class. These nearest points are referred to as support vectors.

To exhibit this in a more understandable way, let be a given data set with labeled training examples:

$(x_1, y_1), ..., (x_n, y_n)$, where $x_i \in R^d$ and $y_i \in (-1, +1)$.

Here, x_i represents the feature vector representation and y_i represents the class label (either negative or positive) of the training subset i.

The optimal hyperplane can be described as:

$w^T x + b = 0$

where w is the vector of weights, x is the input feature vector, and b is the bias term.

The values of w and b must satisfy the following inequalities for all elements of the training subset:

$w^T x_i + b \geq +1$ if $y_i = 1$

$w^T x_i + b \leq -1$ if $y_i = -1$

A SVM model aims to find the values of w and b that separate the data and maximize the margin, which is given by $1/||w||^2$.

The vectors x_i for which $|y_i|(w^T x_i + b) = 1$ are called support vectors. Huang et al. [2018]. This can be seen in the following figure:

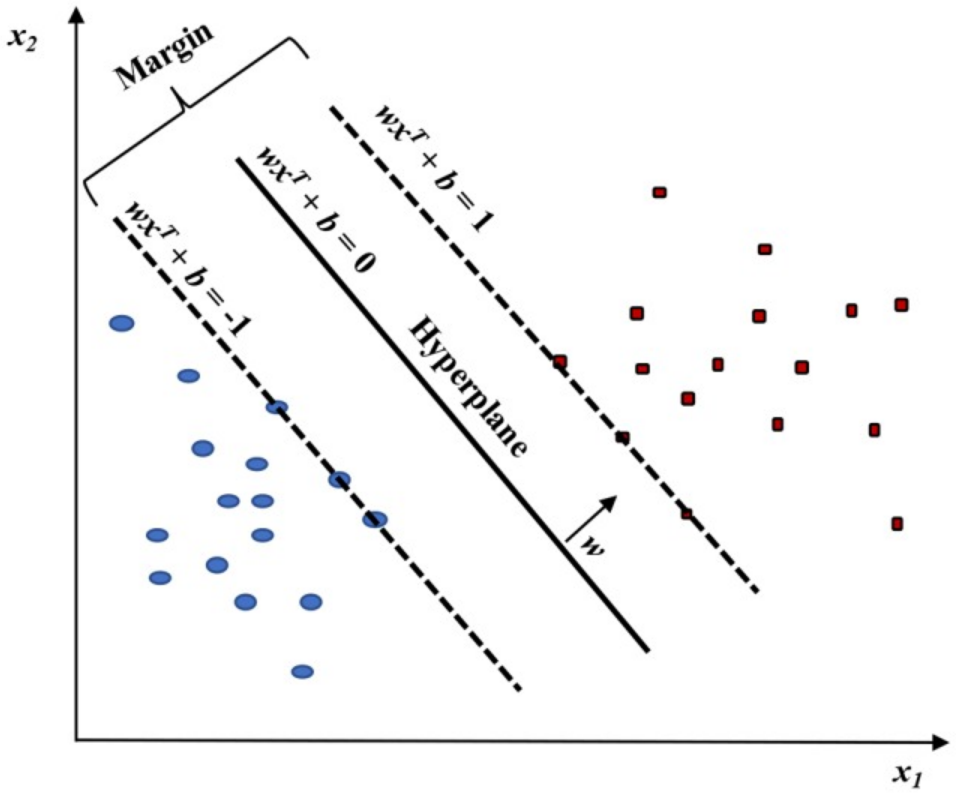

Figure 2. SVM, Linear Kernel

Linear representation of SVM model. Two classes (blue versus red) were classified. By Huang et al. [2018].

Vapnik first introduced the SVM algorithm in 1963 as a means to create a linear classifier. Another application of SVM is the kernel method, which allows us to model non-linear, higher dimensional data. When faced with a non-linear problem, a kernel function can be employed to transform the raw data into a higher dimensional space, thereby making it linear. Essentially,

a kernel function can simplify certain computations that would otherwise require complex calculations in a high dimensional space. Huang et al. [2018]

The kernel functions for larger dimensions are defined as:

$$K(x, y) = \langle f(x), f(y) \rangle \tag{2}$$

where K is the kernel function, x and y are n-dimensional inputs, and f is a function used to map inputs from an n-dimensional space to an m-dimensional space. The dot product between x and y is denoted by $\langle x, y \rangle$. Kernel functions allow us to calculate the scalar product between two data points in a higher dimensional space without explicitly computing the mapping from the input space to the higher dimensional space. Huang et al. [2018]

In many cases, computing the kernel is computationally easy, while computing the inner product of two feature vectors in the high-dimensional space can be difficult. The feature vector can become very large even for simple kernels, and for kernels like the Radial Basis Function (RBF) kernel $K_{RBF}(x, y) = \exp(-\gamma|x - y|^2)$, the corresponding feature vector is infinite dimensional. However, computing the kernel is almost trivial. Huang et al. [2018]

The choice of kernel function can greatly affect the performance of an SVM model, but it is difficult to determine which kernel would work best for a particular pattern recognition problem. The best way to choose the optimal kernel is through experimentation with a variety of "standard" kernel functions. Starting with a simple SVM, we can then test a variety of kernels to determine which is best for the specific problem at hand. Cross-validation can be used in a statistically rigorous fashion to select the optimal kernel function from a fixed set of kernels. Huang et al. [2018]

3.3.1. Implementation of Support Vector Machine Revisiting the approach of the present paper of predicting fully paid and unpaid loans, a SVM model was implemented on the same all data sample from the RF model, with the same proportion of training and test data, 80 and 20 percent, respectively. A linear kernel function was used as an attempt to reduce the computational complexity. The results, as Madaan et al. [2021] found in the research of other papers, SVM accuracy was lower than the RF model one. It will be more explained on the *Results* section.

3.4. Experiment designing

An experiment represents a change in the conditions of the study problem. The studied 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. Benavides [2023]

The addition of the Charged-off amount (*"ChgOffPrinGr"*) variable in early versions of the RF model lead to an accuracy rounding levels of the 99.99 percent, so it is suspicious that the presence of this particular feature may lead to overfitting, which means that the model can not give accurate predictions with new sample of data. The experiment will involve the mentioned feature, 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. If the loan has a certain amount considered uncollectable, it is highly possible that it is labeled as unpaid.

For the experiment, the levels are going to be constructed with quantiles of 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. It is also significant to mention that for this experiment, although the best model was the one ran

with all data, will be run with the sample data as quantiles analysis from all data showed that even in the 80th percentile the Charged Off amount was zero, so the samples would be even more disproportionate.

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 explore if the proportion of paid cases is different among samples. If it is different according to the test, that would imply that the different levels of Charged off amount impacts differently the paid/unpaid cases.

To perform a proportion comparison test to verify that the independent variable affects the dependent variable (by comparing its multiple constructed levels), a chi square test will be run with the "proportionschisquare()" function from scipy.stats Python´s library, which can be used to conduct a chi-square test to compare proportions across k samples. The chi-square test for proportions does not require the distribution of the samples to be normall, but some conditions need to be met. Samples were tested to explore if they followed a normal distribution with Kolmogorov-Smirnov test in python, where our Null Hypothesis indicates to us that the sample is following a particular distribution, in this case, normal distribution. If our pvalue from this test result below our alpha value of significance (0.05 or 95 percent of significance), our Null Hypothesis is rejected. For all samples the test indicated that the pvalue was 0, falling below the 0.05 alpha value, thus, for each sample the Null Hypothesis of sample following a normal distribution was rejected.

Regarding the other conditions, 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, so no impact of the tested variable over the dependent variable. 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:

Table 1. Table of charged off amounts and payment status

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

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.

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 as it is a more complete model compared with DT) 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 of data treatment, so it is not a feasible 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 in Figure 3 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 ten 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

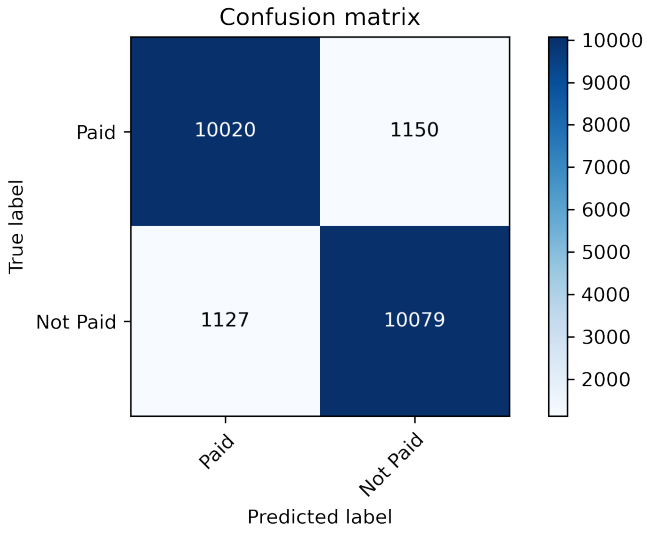

Figure 3. Confusion matrix, RF subset balanced model

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 in Figure 4 the 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 section an experiment is going to be designed to contrast how the particular variable of "Charged off amount" impacts the results of the dependent variable.

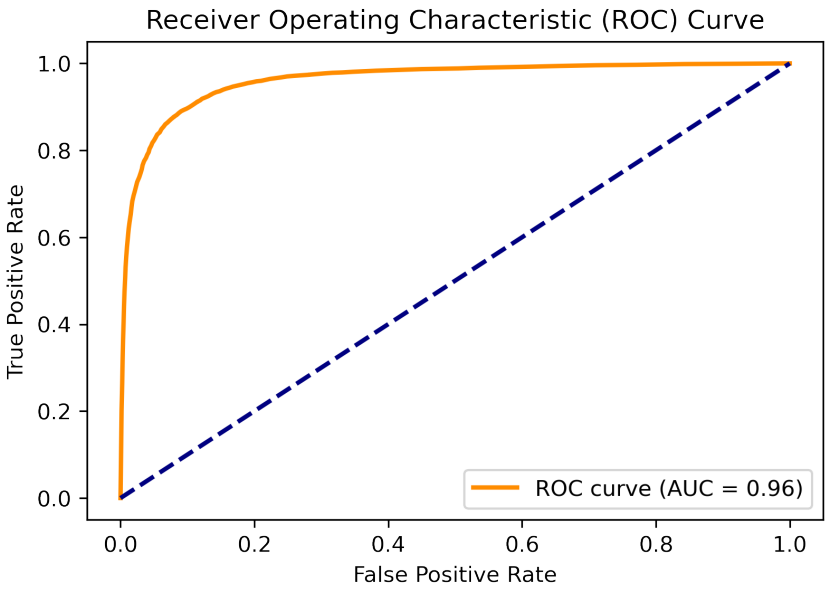

Figure 4. ROC-AUC Curve for RF subset balanced model

4.1. Contrasting with all sample data

There are some justifications to run the Random Forest model over all the sample. First of all, there is an economic sense that the unpaid loans are very infrequent compared with successful paid loans. Banks usually rate the risk profile of each client/company they lend to. Thus, it is more likely that a loan won't be approved to a company with higher risk compared with a company with good credit history and good financial indicators. The other reason is laid on a statistical point of view, where it is wanted to see if the model's effectiveness can be improved. As an accuracy of 89.82 percent was obtained, there may be some gap to close to get a higher percentage of accuracy.

Without further ado, another RF model was run over the all sample data, containing 313 k records, where 257 k records are labeled as successfully paid loans, and 55 k as unpaid (82 percent and 18 percent, respectively rounded); the model was ran with the same parameters: an 80/20 ratio between training and testing split of data, and same number of estimators (100 decision trees). As more records were added, the accuracy index increased to 93.43 percent, and with a Cross Validation technique of 10 folds to test if the model can predict well with new data, a mean accuracy of 93.04 percent was obtained. A improvement in the accuracy of the 3.5 percent was obtained with this iteration. In Figures 5 and 6 a comparison between the two confusion matrix can be seen.

The values from False Positives and False Negatives in both cases may be high, so another additional index was calculated on both results. The precision (also known as positive predictive value) measures the proportion of cases that the model identified as positive that were actually positive. It can be interpreted as the accuracy of the model's positive predictions. A high precision indicates that the model makes fewer false positive errors. For example, in email spam detection, a high precision value would indicate that the model accurately identifies most of the spam emails and does not classify legitimate emails as spam. In the all sample data results, a precision metric of 94.91 percent was obtained, compared with a precision of 89.88 percent of the subsample model.

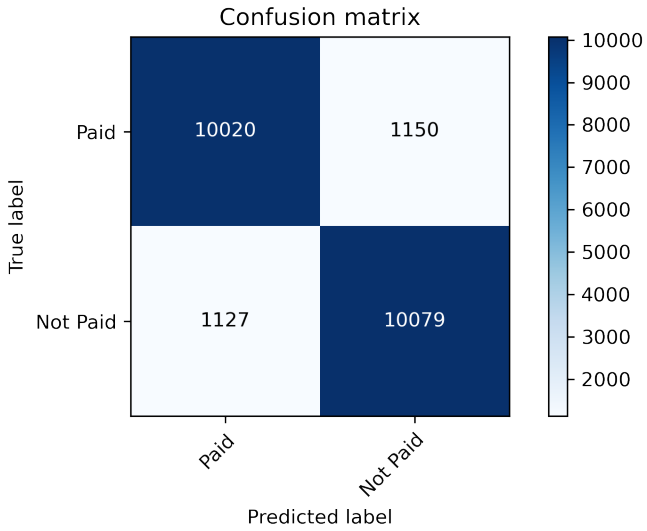

Figure 5. Confusion matrix, RF subset balanced model

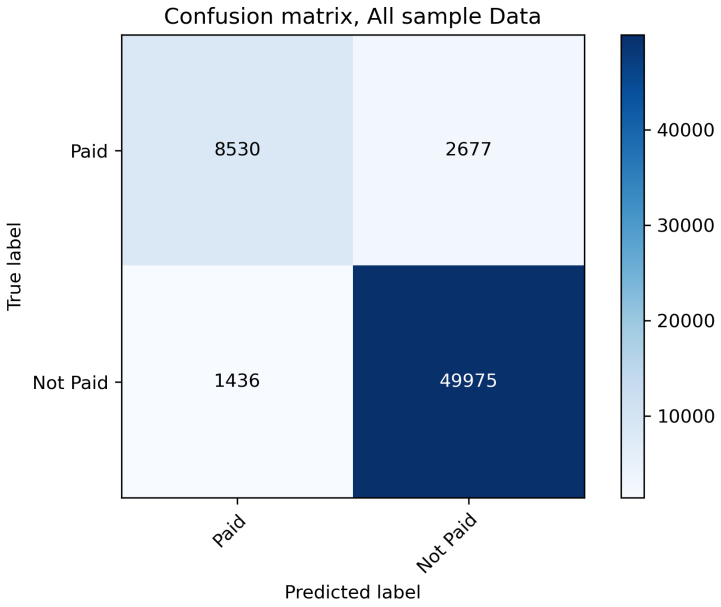

Figure 6. Confusion matrix, RF all sample imbalanced model

Another metric, recall (also known as sensitivity or true positive rate) measures the proportion of actual positive cases that the model correctly identified as positive. It can be interpreted as the ability of the model to identify all positive cases. A high recall indicates that the model is good at identifying positive cases and has a low false negative rate. For example, in medical diagnosis, a high recall value would indicate that the model is able to identify most of the patients with a particular disease. In our example, for the all sample data, an index of 97.20 percent was obtained for the recall, contrasted with the 89.70 percent of the subsample model recall value. Both values, including with the accuracy, indicates that the all sample data model was better than the sub sample data one.

The only thing worth noting left is that the all data sample model consists on a imbalanced data sample. So another metric is also used for this case: the G-mean. The G-mean (also known as geometric mean) is a performance metric that combines the precision and recall of a binary classifier into a single value. It is calculated as the square root of the product of precision and recall, which makes it sensitive to both metrics. Accuracy can be biased towards the majority class, while the G-mean provides a more balanced evaluation of the model's performance across both classes.

The G-mean can be interpreted as the balance between the model's ability to identify positive cases (recall) and its ability to make accurate positive predictions (precision). A high G-mean indicates that the model has both high recall and high precision values, and that it is performing well overall. A low G-mean indicates that the model is either not able to identify all positive cases or is making too many false positive predictions. For our case, in the all data sample model, a G-mean of 96.05 percent was obtained, indicating that the model also performs great even after the fact of imbalanced data.

As a quick, but important notice, is that AUC curves values were the same in both models (all data and sample one), with an AUC index of 0.96. This may be caused because of the effectiveness of the Random Forest model to correctly label both positive and negative cases, and the character of the type of problem being faced: it is expectable to have a great proportion of paid in full loans compared to default loans.

With the addition of this model, the Feature Importance (FI) module of python of the RF models was obtained also, to compare with the results of the sub sample data model. There were no substantially changes, *"Term"* still remained the top characteristic with 54.21 percent of FI, and the other top variables. The unique relatively important change was the Disbursement Gross Amount being more relevant than the SBA Approved Amount, with 7.10 percent of FI. All remain variables stayed the same in terms of order on the FI. It is worth noting that, although at first it was expected that NAIC categorization may help to influence the outcome of the status of the loan, in both iterations the Feature Importances of those variables remained in some of the last places (values very close to zero), so it is possible that they can be disposable for the model.

4.2. Contrasting with Support Vector Machine

As it was reported on the *Methodology* section, results of SVM model were less effective and efficient compared to RF ones. SVM accuracy was lower than the one from the RF model. The Figure 7 shows a confussion matrix protraying this.

An overall accuracy of 87.10 percent was obtained, very much lower than the all sample data RF model of 93.43 percent, so even though SVM models work splitting the hyperplane, a more easy-understandable approach like decision trees can work better. It is also worth to point out that due to the complexity of the SVM algorithm, the duration of the implementation of the code lasted a couple of hours, comparing with the RF that took a duration of less than 10 minutes.

5. Experiment results

Recalling our contingency table contrasting if proportions of paid in full cases vs failed to be paid cases depending on the quantity of charged off amount:

After implementing the chi-square test of proportions, these are the following results: with a level of 2 Degrees of Freedom (three samples minus one), a p value of 0.0 was obtained, and contrasting with our *alpha* of 0.05 (or, 95 percent of confidence), 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

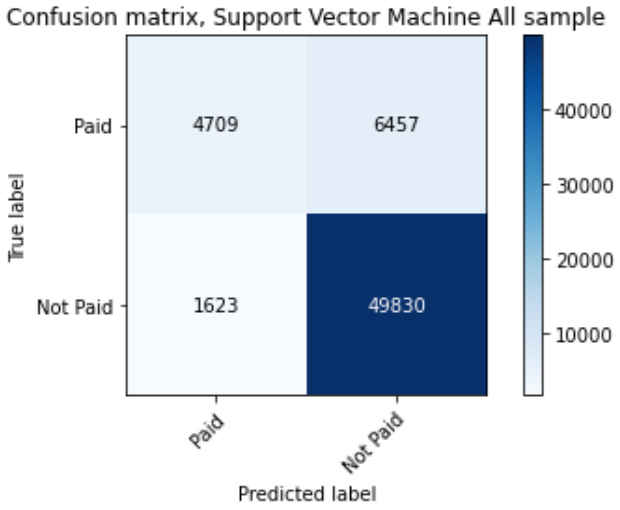

Figure 7. Confusion matrix for SVM all sample data model

Table 2. Table of charged off amounts and payment status

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

amount impacts directly on Paid/Non Paid results. Thus, the level of Charged off amount do change the proportion of fully paid cases.

It is relevant to highlight that this methodology was tried with the all sample data model, unfortunately, when quantile distribution analysis was reviewed, a great proportion of the Charged off amount values were zero. More significant values began to be observed on late quantiles, for instance, the 75th percent quantile. Thus, a partition of the charged off amount records would be more uneven compared with the initial balanced sample data.

6. Conclusion

It is highly noticeable that depending the approach the researcher takes, a new and different insight or result can result. As this paper seeks to develop an adequate model in terms of accuracy but also insights, multiple approaches were compared with other authors and were also contrasted with each other to fulfill this objective.

As Li et al. [2018] studied the same data set as this present paper, some conclusions can come out out of the differences between the approaches. Li et al. [2018] decided to use a logistic regression model with the objective of gaining more insights, while this paper, following the recomendatios of Madaan et al. [2021] to lean on RF models with the objective of maximizing accuracy predictions. This resulted on diverse conclusions. Li et al. [2018] worked with records containing records from both categories of economic cycle of growth and decline. They discovered with a coefficient of their logistic regression that the presence of the economic crisis enhances the probability of not paying a loan. As we wanted to discover insights regarding the dynamic of the loans and the economic cycle, that paper can be contrasted with the approach of the current paper of assuming normal economic cycle of growth, it was discovered, in both *all sample data* and *subset balanced data* RF models, with the *Feature Importance* option of RF

models (which serves as proxy of how much each variable determines the outcome of the model) , that the variables regarding the Term in months that the business have to pay the loan, and the time differences between dates of approval of loans and disbursement of the loans, are the most important variables regarding the process of a payment of a loan. Said in other words, assuming normal economic growth, the more time a business have to organize the dynamic of the payment of the loan, the more probable it is for the business to pay it. The other components that were also determinant for the outcome of this were the ones refering to the quantity of money that could be approved for guarantee for entities like the SBA and the banks, and Number of Employees can also be considered but not so much as this variable contained a FI rounding the 5 percent in both models. At the beginning of the exercise it wasw thought that the NAICS categorization could help to explain better the outcomes, (with the assumption that diverse industries can have better finance ratios compared to another ones), but it was not the case.

Regarding the comparison between the models proposed on this paper, RF models resulted on very similar conclusions regarding the importance of the variables in the models, with the exception of Disbursement Gross Amount being more relevant than the SBA Approved Amount in the *all sample data* model. But, comparing the metrics between the models (89.82 percent accuracy vs 93.04 percent accuracy, respectively), and cross validation technique showing the same trend, can draw some interesting findings. This may indicate that, in general, variables can have similar impacts over the dependent variable of *"paid loans"* across the models, disregarding of number of records and imbalanced data.

It is worth noting that in the development of this paper some variables were left out as it was suspicious that they may lead to an over fitting as it were very relatable to dependent variable. This was checked with the Charged off amount, were it was seen that depending the level of the charged off amount, the proportion of cases of fully paid and defaulted loans were statistically different. Further investigations can be done regarding the relationship between these two variables.

Another conclusion that can come out is that, as economic common sense suggests that in a normal economic cycle of growth, as it is more probable to have businesses with good financial indicators, it is more feasible to have more fully paid loans compared with unpaid loans; so this may be a cause for *all sample data* model to perform better overall contrasted with *subset balanced data*. Even though it was an imbalanced dataset, G-mean technique showed that the imbalanced model performs great with imbalanced data. This may lead to the conclusion that it is possible that the random sampling technique for subset balanced data model may be inadequate and in future works could be extended to a better technique of sampling. Another conclusion that can be drawn is that, even though some ML algorithms have high mathematical procedures like the Support Vector Machines, Random Forest models adjusts better to this kind of problematic compared with the SVM ones. It is worth mention again that the approach the researcher will take determine the insight. In this paper accuracy of predictions were the main task but also it was important to get insights. The present paper can extend to clustering analysis to seek for patterns, and also to explore the model including the whole batch with the objective to contrast growth and crisis cycles. Further investigations may take place regarding some new variables concerning financial ratios, economic indicators and so on, that may lead to even higher accuracy metrics and new insights in regards to the science of credit scoring.

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-/ c4b2c8cc9feb. Accessed: February 26, 2023.

Shujun Huang, Ni Cai, Pedro Penzuti Pacheco, Sebestian Narrandes, Yuliang Wang, and Wei Xu. Applications of support vector machine (svm) learning in cancer genomics. *Cancer genomics & proteomics*, 15(1):41–51, 2018.

Alberto Benavides. Aprendizaje automático, 03 2023. URL `https://github.com/albertobenavides/aprendizaje_autom`.