CIENCIA *ergo-sum*
Universidad Autónoma del Estado de México
ciencia.ergosum@yahoo.com.mx
E-ISSN: 2395-8782

Portfolio recommendations to improve risk of default in microfinance

Simonin, I., Brooks, M. & Nieto-Barajas, L. (2021). Portfolio recommendations to improve risk of default in microfinance. CIENCIA *ergo-sum*, *28*(1). https://doi.org/10.30878/ces.v28n1a6

Portfolio recommendations to improve risk of default in microfinance

Recomendaciones de cartera para mejorar el riesgo de incumplimiento en las microfinanzas

Irving Simonin
Universidad Nacional Autónoma de México, México
isimoninw@gmail.com
https://orcid.org/0000-0002-6298-2359

Marc Brooks
Duke University, Estados Unidos de América
mgbrooks208@gmail.com
https://orcid.org/0000-0002-4219-6422

Luis Nieto-Barajas
Instituto Tecnológico Autónomo de México, México
lnieto@itam.mx
https://orcid.org/0000-0002-0859-7679

Recepción: 22 de octubre de 2019
Aprobación: 20 de marzo de 2020

Abstract

This article presents an exciting application of machine learning for loan origination in microfinance. Microfinance targets people who cannot build a credit history and therefore cannot access loans from banks or other financial institutions. We use data from a Mexican microfinance company that operates in several regions throughout the country. The objective is to guide intermediate lenders to choose their clients and achieve a lowerr credit default risk. We use several statistical models such as principal component analysis, clustering analysis and a regression tree. We obtain, as a result, a series of recommendations based on the characteristics of the clients.

Keywords: Clustering analysis, machine learning, microfinance, risk of default, principal components, regression tree.

Resumen

Se presenta una interesante aplicación de aprendizaje de máquina en la originación de créditos en microfinanzas. El objetivo de microfinanzas son las personas que no pueden construir un historial crediticio y, en consecuencia, no pueden acceder a préstamos de bancos u otras instituciones financieras. Usamos datos de una compañía microfinanciera mexicana que opera en varias regiones del país. De igual modo, se pretende guiar a prestamistas intermediarios para escoger sus clientes y alcanzar un menor riesgo de crédito. Usamos varios modelos estadísticos como análisis de componentes principales, análisis de grupos y árboles de regresión. Obtenemos, como resultado, una serie de recomendaciones basadas en las características de los clientes.

Palabras clave: análisis de grupos, aprendizaje de máquina, árboles de regresión, componentes principales, microfinanzas, riesgo de crédito.

Introduction

Extreme poverty is an unfortunate cycle, where low-income individuals lack the financial capacity to lift themselves out of poverty with diminished capacity to start a business, invest in their professional abilities, or other profitable activities. Microfinance attempts to break this cycle by providing small loans to low-income individuals (Condusef, 2014). With these loans, individuals have the financial basis to engage in profitable pursuits with a low risk of aggregating a large debt. In Mexico, the focus of the microfinance companies has been vulnerable sectors of the population, mainly rural areas and specially women; those who lack a way to build a historical credit score and in consequence cannot access loans from banks and other financial institutions. For an overview of the history of microfinance, including Mexico, and a list of the most common statistical techniques used for credit risk measurement, see Lara-Rubio (2010).

This article presents an exciting application of machine learning for loan origination in microfinance. We use data from a Mexican microfinance company that operates in several regions throughout the country. Their direct clients, which are mostly female, have a revolving loan that they can then distribute as smaller loans to their own personally selected client base. We will refer to the primary clients, the intermediary lenders, as "lenders", and their clients as the "final clients". The lenders can have up to a 250 000 MXN cap on their credit limit and can lend a maximum of 6 000 MXN to each client. There is no restriction on the number of final clients. Loans are paid off in fixed payments each fortnight (biweekly), which should be collected by the lenders. Lenders that are unable to sufficiently collect their payments must pay what is owed themselves. A lender's own credit limit is determined by their ability to make punctual and complete payments as well as the consistent use of the credit limit provided to them. Consistently incomplete and unpunctual payments naturally lead to the termination of their account. As a practitioner of microfinance, the company's goal is not only to reduce its losses from credit default but also help their clients achieve financial success and economic independence. The final clients that lenders select are critical to achieving this goal. Our task is to use statistical models to guide lenders as they develop their portfolio of clients.

The content of the rest of the paper is as follows: Section 1 describes the available information. Section 2 contains the data analysis based on principal components analysis, a hierarchical clustering technique, and a regression tree. Section 3 shows the results and the format of the model's recommendation, and the last section concludes with a discussion.

1. Available information

While there is a significant amount of information on the lenders, the data on the final clients is minimal. More specifically, since the final client payment is collected directly by lenders, and the payment to the company comes as a single biweekly payment by the lender, there are no records of final client behaviour. The most relevant data describing the final clients is demographic data and general descriptive variables of the specific loans, such as the amortization period as well as the amount of the installments. "Amortization period" refers to the number of fortnights the client is given to pay off the loan. The installment amount is the amount of the loan divided by its amortization period. In addition, there exists a score that quantifies the uncertainty of credit default for each lender, which we refer to as "credit default risk". Using this data, we construct an ideal client/loan profile for a specific lender. This materializes into two types of recommendations that are made:

a) General guidance on the client profiles and loans a lender should look for (e. g. young females ages 28-36 with an amortization period of 6 fortnights).

b) Given a particular client profile (e. g., a young single male) they can determine the type of loan that should be offered, specifically the loan amount and amortization period.

Portfolios can change at any point in time as a lender can give new loans to new clients or even new loans to current clients. We summarize the portfolios by computing the weighted mean and weighted standard deviation of the client demographic variables. The weight is the client's installment amount divided by their lender's total installment amount for that fortnight. Other important explanatory variables are the age, gender, and civil status of the lender herself, although these are not used in the recommendation because they can not be modified. Therefore, the lenders' information consists of multivariate time series observed every fortnight. In summary, the key response variable is risk of default, and the explanatory variables are divided into two types, three lender features: age, marital status and gender; and 11 portfolio features: the number of final clients, the number of loans, mean percentage of married final clients, plus the mean and standard deviation of: amortization period, biweekly expected payment, percentage of male final clients and age of the final clients. We use data from 2018 for this study and restrict the training data to lenders that have more than one client and have been borrowing from the company for more than 3 months. Therefore, we are left with 684 000 data points, which correspond to 36 986 unique lenders and 23 fortnights.

An example of the time series of a lender is shown in Figure 1. The idea is to characterize the conditional distribution of the response given the explanatory variables, with the final aim of advising a lender to change features of her clients to reduce her risk.

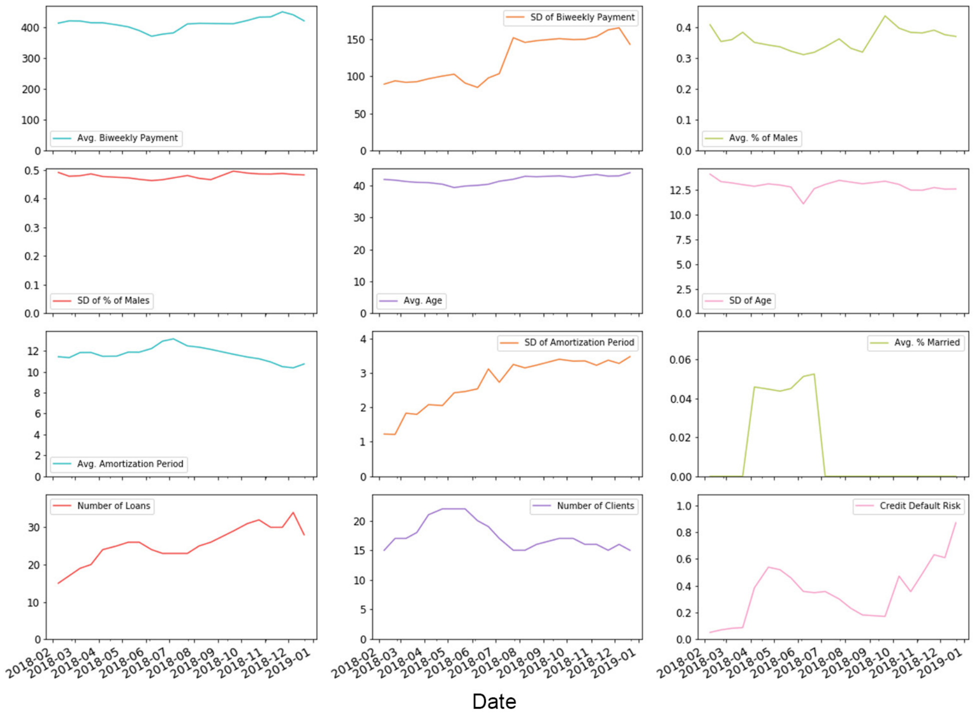

FIGURE 1

Time Series of a lender's portfolio characteristics and her credit default risk. The lender was selected randomly from lenders with more than 0.6 in credit default risk (high risk)

Source: Self-created.

2. Data analysis

The company operates in over 100 zones across Mexico with different socioeconomic and geographical factors. Available data corresponds to only 89 zones. Instead of performing an analysis for each of these zones, we cluster them based on their lender's explanatory variables over time. We first define variables at zone level by calculating the mean, standard deviation, and median for the 14 variables for all lenders in a zone for each of the fortnights. We then reduce the dimensionality of the data and produce a single time series indicator per zone. For this, we use principal component analysis (Lever *et al.*, 2017) and kept only the first component which explained 24% of the total variability. Since our objective is not to replace the original variables, but to cluster the zones, we believe such variability explained by the first component is reasonable. With the first principal component we implemented hierarchical clustering using Ward's minimum variance method (Murtagh & Legendre, 2014) to form groups of zones whose trends are more similar over time. We preferred a hierarchical clustering over the most common method of k-means, because the output given in a dendrogram allows us to select the best number of clusters. Additionally, we chose the method of Ward, because it is the only hierarchical method based on a sum of squares criterion, producing groups that minimize within-group dispersion.

This leaves us with 16 groups (17 including zones with less than a year of operation) and the number of zones per group ranges from 1-10. These final clusters are shown in Figure 2. Here we observe that the behaviour of the zones inside the same group are very similar and distinct to those in different groups.

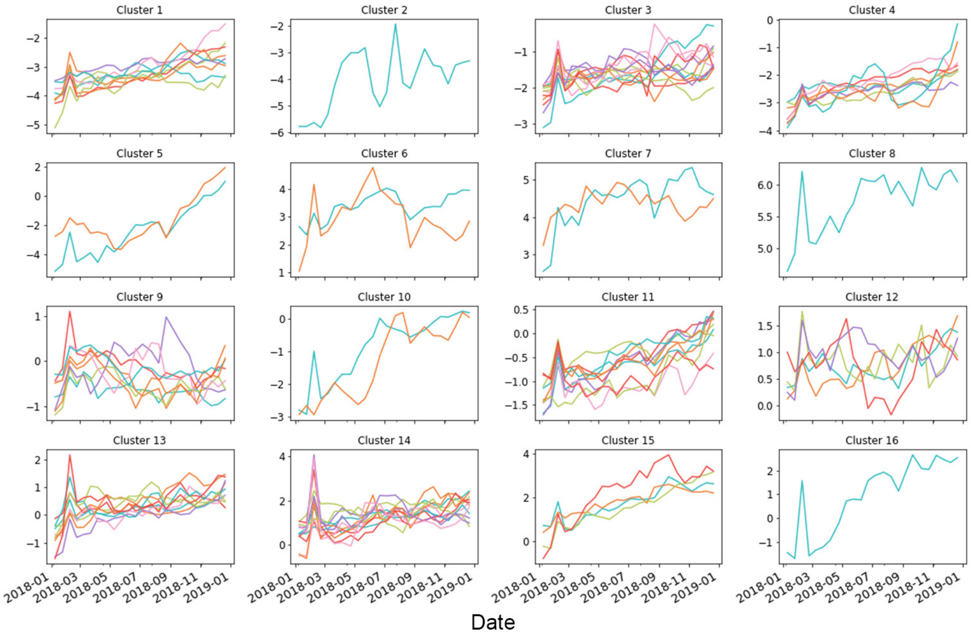

FIGURE 2

Time series of the first principal component for each zone grouped by assigned cluster

Source: Self-created.

As mentioned, the goal of this project is to provide sensible and understandable recommendations for the lenders as a way to guide their own client selection. According to Schreiner (2002), one of the most effective credit scoring methods is the scorecards which is based on trees. Differing from random forests (Breiman, 2001), where predictions are usually better than in a single tree, interpretation of how explanatory variables interact to produce a specific response is not possible to grasp. In particular, regression trees (Breiman *et al*., 1984; Loh, 2008) create a partition of the training sample, based on easily interpreted partition rules of the explanatory variables. For each cluster we train a regression tree using the risk as dependent variable and the set of lenders characteristics as explanatory variables. The tree is constructed through a sequence of partition (decision) rules that continue to split the data into "purer" (more homogeneous) subsets until arriving at a final node. These rules are defined by finding a threshold on the series of features. An example is given in Figure 3, where the first node is split according to whether the mean amortization period is less than 12.625 fortnights or not.

We created several trees at different depths, using the mean absolute error as splitting criteria, and selected the one with the smallest prediction error and terminal nodes with at least ten lenders. Each final tree has a depth of 13 levels. As mentioned before, a decision tree can be seen as a partition of the feature space into homogeneous classes (nodes) of the response variable, we use the partition rules to assist us in creating recommendations for the lenders. In our case we are performing regression, so the prediction is the median of the response variable for all the individuals assigned to a particular terminal node (or leaf), in the training sample. Depending on the lender's specific feature values, we can find the path of each lender down the tree. Therefore, one could suggest adjusting by specific variables and assigning the lender to a different leaf with lower risk. This forms the basis of our recommendations.

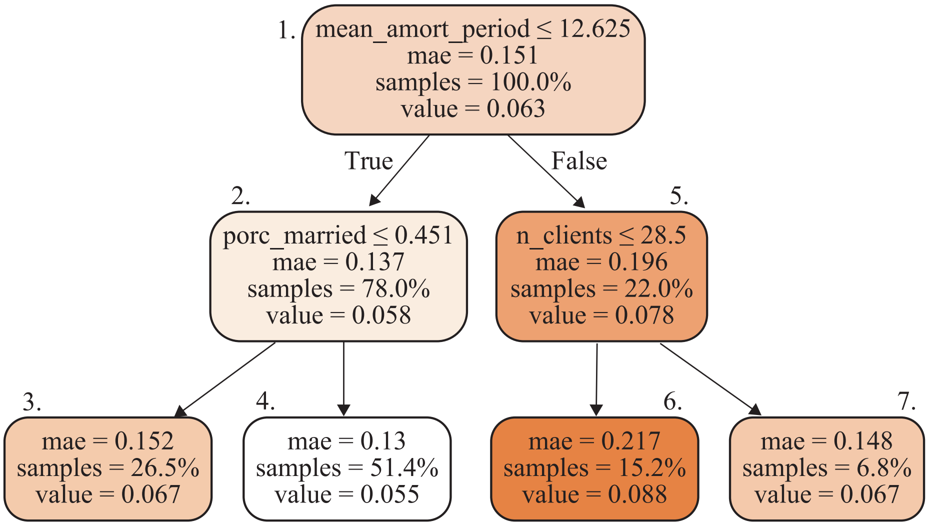

FIGURE 3

Regression tree with a depth of two levels and four terminal nodes (leaves)

Source: Self-created.

There are often numerous potential leaves that can be used as the base for a recommendation. This is where the idea of sensible recommendation plays an important role. An ideal recommendation is one that improves the risk without requiring extensive changes to their client / loan portfolio. We also note that there could be implausible recommendations, such as changes in lenders' gender, age or marital status. To avoid this, we marginalize the splits defined by any of these variables.

Finally, to find the best recommendation we derive a measure of similarity between the different potential leaves. Each leaf can be represented as a unique vector of zeros and ones of length given by the total number of nodes in the tree. All nodes that belong to the path of a leaf will have a value of one, and value of zero otherwise. These are sparse vectors that contain many zeros and at most 14 ones. In the example of Figure3, each vector would have length seven as a result of the number of nodes used in the tree. For instance, the terminal node furthest to the left would be represented as. The similarity is based on the cosine similarity metric, which is defined as the normalized Euclidean dot product of two vectors. This metric is commonly interpreted as the cosine of the angle between the two vectors. A cosine value of 1 indicates the vectors are parallel while a value 0 indicates they are perpendicular. The closer the value is to 1, the more the two vectors overlap.

For each leaf, we sort the other leaves by two criteria. First, the cosine similarity and second, the predicted risk. In this way, we are looking at the leaves that are most similar to the leaf a particular lender has been assigned to and then we select the leaf with the lowest predicted risk. It is a method to obtain recommendations that are simple and understandable for the lender. An example for the first 3 of these nodes for a lender with risk 0.88 are presented in Table 1. Each node returns the feature that differs from our lender's feature values and the threshold the lender would need to meet to be assigned to that node.

TABLE 1

Sorted recommendations for a lender with risk value of 0.88.

Feature	Sign	Threshold	Original Value
Leaf 1: Median risk = 0.033, Cosine similarity = 0.917			
Mean payment (biweekly)	>	582.5	550.0
Leaf 2: Median risk = 0.026, Cosine similarity = 0.772			
SD of amortization period	≦	3.04	3.38
SD of % male clients	≦	0.498	0.499
Leaf 3: Median risk = 0.032, Cosine similarity = 0.772			
SD of amortization period	≦	3.04	3.38
Mean % clientes married	≦	0.45	1.0

Source: Self-created.

3. Results

At this point we have developed methods for selecting potential leaves that will form the basis of our recommendations. Rather than giving a precise action a lender should take, we suggest general guidelines they should adopt. This derives from a main assumption: since all loans are short to mid-term, lenders have a rotation of clients. By using these guidelines when they establish the following inflow of clients, little by little their portfolio will shift and eventually comply with partition rules of the leaf.

Let us consider the example of leaf 1 in Table 1. In order for the lender to move to a leaf with lower risk, she has to increase their average clients' biweekly payments above 582.5. This can be done by generating loans of larger amounts or by decreasing the amortization periods they have been using. Examine now the 2nd leaf. The first partition rule indicates the lender should place loans of no more than +/- 3.04 of her current mean amortization period. In our example the lender's mean amortization period is 18.68, therefore we have 15.64 and 21.72 as upper and lower bounds for intervals to recommend. Since the maximum length of amortization period is 20, we recommend that the client issues loans with an amortization period between 16 and 20 fortnights.

Now consider the second partition rule from leaf 2. Again, it states that the lender should decrease the SD of their percentage. In this case we must translate the SD of a binary variable. Our lender's clients are 52% male thus the easiest way to decrease the SD would be to increase her percentage of male clients. Using the definition of the SD of a Bernoulli, to derive a new interval for the percentage of men. Values between 55% and 100% would sufficiently decrease the SD. Finally, In the second partition rule of leaf 3 the lender should decrease her percentage of married clients from 100% to lower than 45%. The recommendation would simply advocate a portfolio with more single clients.

It is important to remember that these are general guidelines we recommend a lender to adopt. They are broad intervals that give the lender's discretion in their decision making. As they adjust to these guidelines, they will develop a portfolio that corresponds to lower risk of credit default and in turn improve their own credit limit with the company.

Conclusion

In this article we managed to successfully implement statistical models to help a microfinance company to reduce financial risk. All data processing was done in Python, which is a powerful tool to deal with large datasets. For the principal component's analysis and regression tree, the *scikit-learn* module was used. Data analysis followed these steps: cleaning the data, defining variables, dimensionality reduction, creation of groups, constructing a regression tree and finally converting the model's results into user-readable output. The whole process can be summarized in the diagram of Figure 4.

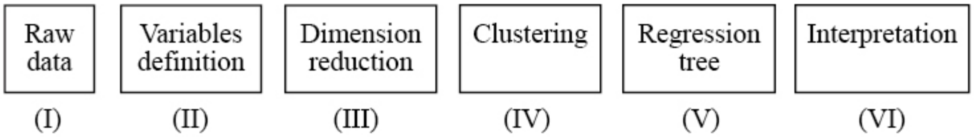

FIGURE 4
Data analysis process. Steps (I) to (VI)
Source: Self-created.

Working with real data is always a challenge and this was not an exception. Open questions still remain, such as updating the analysis, and linking subsequent recommendations with previous ones perhaps by defining a dynamic tree constructed at different time points.

Acknowledgements

This work was developed while the first two authors had a professional affiliation with Datank AI, Mexico.

References

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, C. A.: Wadsworth & Brooks.

Breiman, L. (2001). *Machine Learning*, *45*, 5. 10.1023/A:1010933404324

Condusef. (2014). Microcréditos, el costo de contratarlos. *Proteja su dinero*. Retrieved from https://www.condusef.gob.mx/Revista/index.php/credito/personal/404-microcreditos

Lara-Rubio, J. (2010). *La gestión del riesgo de crédito en las instituciones de microfinanzas*. Retrieved from https://hera.ugr.es/tesisugr/18892656.pdf

Lever, J., Krzywinski, M., & Altman, N. (2017). Principal component analysis. *Nature Methods, 14*, 641-642. https://doi.org/10.1038/nmeth.4346

Loh, W.-Y. (2008). Classification and regression tree methods. In *Encyclopedia of Statistics in Quality and Reliability* (pp. 315-323). Wiley.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*, *31*, 274-295.

Schreiner, M. (2002). Scoring: The Next Breakthrough in Microcredit? *Technical report*. Retrieved from http://www.microfinance.com/English/Papers/Scoring_Breakthrough.pdf