Matheus Andrade dos Reis

Modelo Preditivo de Risco de Crédito para Cooperativas de Agronegócio

Brazil

2022

Matheus Andrade dos Reis

Modelo Preditivo de Risco de Crédito para Cooperativas de Agronegócio

Dissertation presented to the Professional Master's Course in Economics, University of Brasília, as a partial requirement for obtaining the Master's degree in Economics

Universidade de Brasília - UnB

Faculdade de Administração Contabilidade e Economia - FACE

Departamento de Economia - ECO

Programa de Pós-Graduação

Supervisor: Daniel Oliveira Cajueiro

Brazil

2022

Matheus Andrade dos Reis
Credit Risk Predictive Model for Agribusiness Cooperatives/ Matheus Andrade dos Reis. – Brazil, 2022
35p. : il.; 30 cm.

Supervisor: Daniel Oliveira Cajueiro

Dissertation (Masters) – Universidade de Brasília - UnB
Faculdade de Administração Contabilidade e Economia - FACE
Departamento de Economia - ECO
Programa de Pós-Graduação, 2022.

1. Credit Risk. 2. Machine Learning. 3. Agribusiness Cooperatives. II. Universidade de Brasília. III. Faculdade de Administração, Contabilidade e Economia - FACE. IV. Departamento de Economia IV. Credit Risk Predictive Model for Agribusiness Cooperatives

Matheus Andrade dos Reis

Modelo Preditivo de Risco de Crédito para Cooperativas de Agronegócio

Dissertation presented to the Professional Master's Course in Economics, University of Brasília, as a partial requirement for obtaining the Master's degree in Economics

Approved. Brazil, May 30th, 2022:

Daniel Oliveira Cajueiro
Chair

Herbert Kimura
Member

Pablo José Campos de Carvalho
Member

Brazil
2022

To my family and friends,
who have been always there for me.

Acknowledgements

I would like to thank Jehovah and my family, who gave me the gift of life and all the support I needed to get here.

I would also like to express my gratitude to my primary supervisor, Daniel Cajueiro, who guided me throughout this project.

I wish to acknowledge the help provided by many workmates from Caixa Economica Federal, who not only helped with my data for this study but also understood the time I had to invest on this project.

I would also like to show my deep appreciation to all the friends who have been with me during this important stage of my life.

"The heart of the understanding one acquires knowledge
And the ear of the wise seeks to find knowledge."
(New World Translation of the Holy Scriptures, Proverbs 18:15)

Abstract

The use of Machine Learning techniques in risk management is an increasingly common practice in corporations. Banks, in particular, use these techniques in their credit risk analysis processes. This research evaluates the use of common Machine Learning models to assess the credit risk of a specific public: agribusiness cooperatives. Various models consolidated in the market were tested and we verified that, for this purpose, two models stood out: Gradient Boosting and Random Forest. The models revealed that the variables related to the economic-financial situation of the cooperatives, measured by financial ratios, are more relevant than other information obtained in several researches, such as the behavior in other financial transactions and the administrative capacity of the cooperative.

Palavras-chave: Credit risk, Machine Learning, Agribusiness Cooperatives.

Resumo

O uso de técnicas de Machine Learning na gestão de riscos é uma prática cada vez mais comum nas corporações. Em especial os bancos tem utilizado essas técnicas em seus processos de análise de risco de crédito. Este trabalho avalia o uso de modelos comuns de Machine Learning para avaliar o risco de crédito de um público específico: cooperativas de agronegócios. Foram testados diversos modelos consolidados no mercado e verificamos que, para este propósito, dois modelos se destacaram: Gradient Boosting e Random Forest. Os modelos revelaram que as variáveis relativas à situação econômico-financeira das cooperativas, medidas por meio de indicadores financeiros, são mais relevantes do que outras informações obtidas em pesquisas diversas, como o comportamento em outras transações financeiras e a capacidade adminsitrativa da cooperativa.

Keywords: Risco de crédito, Machine Learning, Cooperativas de agronegócios.

List of Tables

List of abbreviations and acronyms

ML	Machine Learning
OLS	Ordinary Least Squares
SVM	Support Vector Machine
kNN	k-Nearest Neighbors
EN	Elastic Net
RF	Random Forest
GB	Gradient Boosting
MAE	Mean absolute error
MSE	Mean squared error
MAPE	kMean absolute percentage error
RMSE	Root mean squared error
Ebitda	Earns before interest, taxes, depreciation and amortization

List of symbols

α	Lower case Greek letter alpha
ρ	Lower case Greek letter rho
ζ	Lower case Greek letter zeta
ϕ	Lower case Greek letter phi
ε	Lower case Greek letter epsilon
Θ	Upper case Greek letter Theta
Σ	Upper case Greek letter Sigma

Contents

1 Introduction

Financial institutions, as fundraisers and grantors, are subject, among other risks, to credit risk. The assessment of this specific risk, in several forms, allows institutions to have better conditions to decide whether to grant or not the credit to a customer. In this scenario, different customers are not necessarily evaluated in the same way, acknowledging that it is not the same factors that allow us to assess the risk associated with the credit granted to each of these customers. Following this logic, companies with different structures, activities and objectives need to have their specificities taken into consideration in the credit risk assessment. An example of a customer that presents many differences in relation to others are the farmers cooperatives.

Even though agribusiness cooperatives resemble other business, they also differ in many ways. For example, a cooperative's purpose, its ownership and control, and how benefits are distributed are not the same as in a regular company. These differences may affect many features that are regularly used in credit risk assessments, so it seems reasonable to it seems reasonable to evaluate these cooperatives with a specific focus, and not as other companies in general.

The present research aims to develop a predictive model using supervised learning techniques to assess the credit risk of farmers cooperatives. We use variables usually adopted by the market to evaluate credit risk, some specific information for this kind of customers and the final score of assessments carried out in a fundamentalist model for this public. The objective of developing the new model is to allow the users of this new model, who could be banks managers, for example, to proceed with these evaluations demanding less time and information from customers.

Although the work of knowledgeable business analysts and generally accepted fundamentalist techniques are appropriate, these assessments are often based on measures that adjust slowly over time, so that some information taken into consideration are not always relevant to the tactical decision in risk management (KHANDANI; KIM; LO, 2010). Based on these observations, the opportunity arises to improve the credit risk assessment process for agribusiness cooperatives, by using a supervised learning model to build an algorithm that can assess the customer's credit score with less information and faster than the original model.

When comparing alternative options to an individualized analysis of borrowers, Aniceto, Barboza and Kimura (2020) compared different techniques of machine learning and traditional models based on logistic regressions, concluding that the algorithms of the first option, on average, presented more satisfactory results. Also, according to Addo,

Guegan e Hassani (2018), data science approaches such as machine learning and deep learning models play a significant role in modeling credit risk. The authors also conclude that it is important to consider a variety of models that best fit the available data and business problem. Also on machine learning, Leo, Sharma and Maddulety (2019) highlights that, although it can be used to manage other risks, many studies focus on credit risk models, generally related to credit scoring, and that there is room for application in other problems in this same field. Therefore, this study uses the data obtained to verify which of the already known models will bring better results for this specific purpose.

Supporting and complementing the advantages of such techniques for measuring credit risk, Bussmann, Giudici, Marinelli and Papenbrock (2021) mention that they can show a non-linear relationship with the financial information observed in past statements, which can increase the accuracy of models.

Regarding the most used techniques and their adherence to the data found, Breeden (2020) compared 6 models for credit scoring: logistic regression, SVM (Support Vector Machine), kNN (k-Nearest Neighbors), decision trees, neural networks and ensembles. Considering the results of this study and the low number of agribusiness cooperatives that have fundamentalist credit risk assessment in the same financial institution, kNN may seem to be an appropriate model for the problem to be researched, given the small volume of data available, that is, the evaluations already carried out and the results obtained.

In another perspective, Shoumo et al (2019) compared SVM with three other supervised learning models and, for the data set available in that case, which is a peer-to-peer credit institution, found that this model proved to be more adequate. Although there are differences between these data and those obtained from agribusiness cooperatives, the study concluded that the referred model is an interesting approach to identify bad debtors in an economic context with constant variations, which is also a common reality for many agricultural producers.

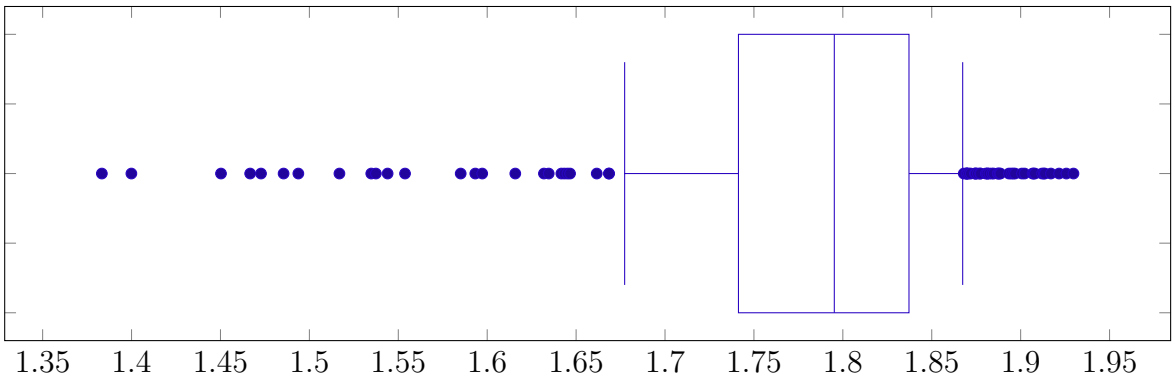
There were found no studies that make applications of machine learning models specifically for the purpose of this study. Therefore, existing research confirms the applicability of these techniques to the problem to be addressed in this paper, so that the development of this model aims to meet a need noticed by financial institutions that seek to improve the process of granting credit to agribusiness cooperatives.

The structure of this dissertation is divided into the following four sections. The next below will describe the data obtained and the treatment given to it so that they could be used in this study. Next, we will briefly describe the methods chosen to find the most suitable Machine Learning model for this case. After that, the results are presented and commented. Finally, we include an objective conclusion about our findings.

2 Data

The dependent variable, which we are trying to predict, is the score of agribusiness cooperatives evaluated in a fundamentalist model of credit risk used by a large Brazilian bank between 2017 and 2022. This score, that ranges from 0 to 100, was used on a log basis in this study, so as to reduce the distance between the samples and make it more normalized [1].

Scores Distribution



In order to predict the scores mentioned above, we collected data from the agribusiness cooperatives evaluated. Financial statements from the past three years were used, so they tend to comprehend the reality in different production cycles. The evaluated cooperatives were from different regions of Brazil. In total, we had 249 cooperatives.

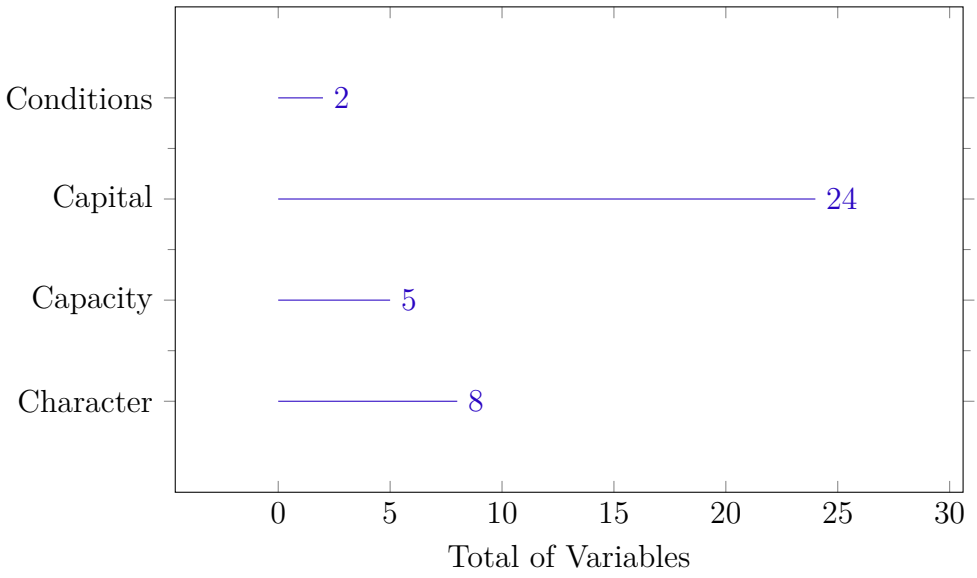
An important part of credit risk assessment is to verify how customers have behaved with creditors and suppliers in their previous transactions. In this way, common qualitative information for this type of analysis was obtained. However, bankruptcy prediction is becoming more based on quantitative than qualitative aspects in recent years (CHOW, 2017). For this reason, we used information from the cooperatives financial statements and calculated financial ratios for the three years prior to the original assessment, which result in the score that we want to predict. We selected eight financial ratios that are regularly used in credit risk assessments, and were commented by Beaver (BEAVER, 1966).

We grouped all the independent variables into four groups, according to the classification of the 5 C's of credit (character, capacity, collateral and conditions). However, since these assessments are limited to measuring the creditworthiness of a customer before a specific evaluation for the requested credit loan, the C of collateral, which refers to the

[1] The target we are trying to predict here is the score in the assessment made by a human analyst prior to granting credit, and not necessarily related to the company's status at the end of the loan term. Eventual mistakes made by the analyst are reflected on the final score.

guarantees offered in the loan was not considered. The four groups will be discussed in sequence.

The distribution of the 39 variables that we kept in our database and their distribution into the mentioned categories can be seen in the figure below[2]:



2.1 Character: Registration information

Customer registration data were some of the variables used in the evaluations. These data was available in the financial institution's records and other specialized firms which provide such information, such as the billing history of cooperatives and complaints about previous default. Other information used include quantity and recency of complaints from suppliers, were also obtained in cadastral surveys. The customer's relationship time with the financial institution and the historical account balance were also considered.

The values of any debts that may not have been paid were weighted by the total sales of the cooperative, allowing to properly compare the debts of customers of different sizes.

2.2 Capacity: Management style

In order to verify the cooperative's ability to keep generating cash and paying off its debts, it was necessary to investigate more deeply about each customer's management style and business strategy. This information might not be easily obtainable. Nevertheless, since the public of these assessments are big farmer‘s cooperatives, it is possible to find useful information in the regional news and on the cooperatives' websites. This way we could gather some information about the cooperative operates and its directors. Basically,

[2] The list of all variables is available on Annex A.

we analyze whether the cooperatives have a clear action strategy, and how they work to finance their members, when this is necessary.

Other variables associated to this group could be obtained through observing the financial statements. We looked at the results volatility throughout the three years that we were analyzing. We also looked for specific balance sheet accounts to check for signs that the cooperative hedges for changes in the prices of its products.

We also put in this category an aspect that came from the cooperative registration in the original database, which is the number of months the cooperative has been operating.

2.3 Conditions: Economic scenario

This category groups together variables that consider the economic situation of the various agribusiness sectors, as well as the cooperatives' ability to deal with these variations. Among the evaluated cooperatives, many work with the production of corn, soy, coffee, dairy, and others. Therefore, they face different scenarios, according to the cycle of their products.

The short-term risk classes of each type of production were obtained through an external agency that studies economic scenarios. In some cases, information from other sources, such as producer associations, was also considered for this classification.

In this category there is also a variable that came from a search through the cooperative's website or, when available, from the most recent income statement, which is the diversification of products sold. We try to check if the dependence of only one or two products would increase the credit risk.

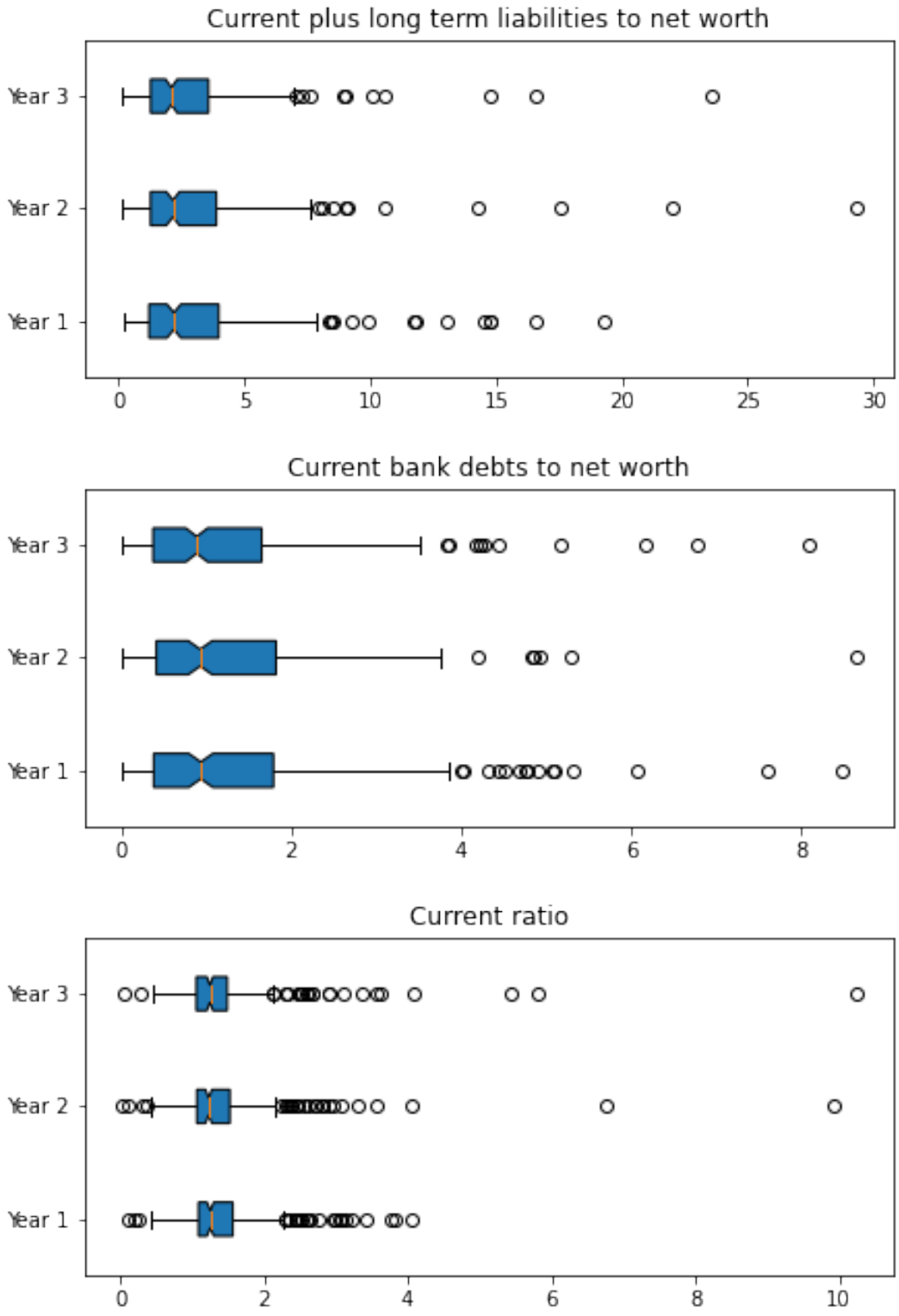
2.4 Capital: Economic and financial ratios

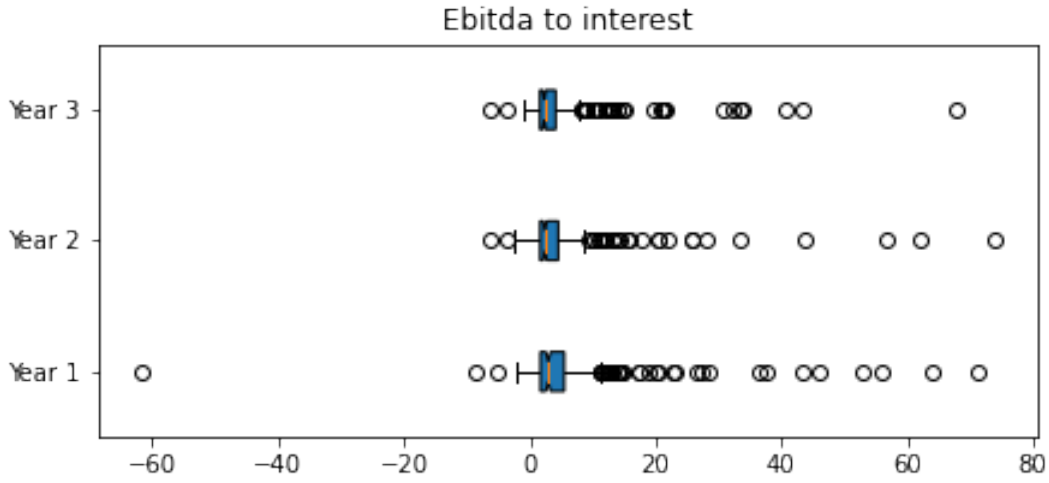
As is usually done in company valuations, an important part of the credit risk assessment is the ratios computed from the financial statements. Eight different ratios were calculated for the three years prior to the analysis.

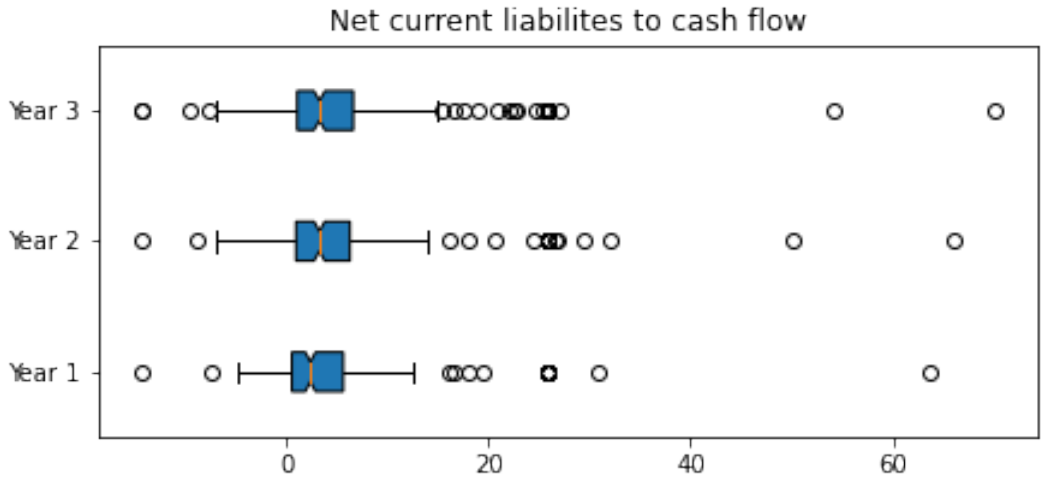
Economic and financial ratios used here aim to make the understanding of various aspects of cooperatives clearer, such as the degree of leverage, financial liquidity, operating cash flow and the profitability of their activities. All this information is used as a measure of the borrower's financial health, which means a lot to the institution that intends to grant the funds. In this sense, the same calculation parameters were used as those usually performed in the market.

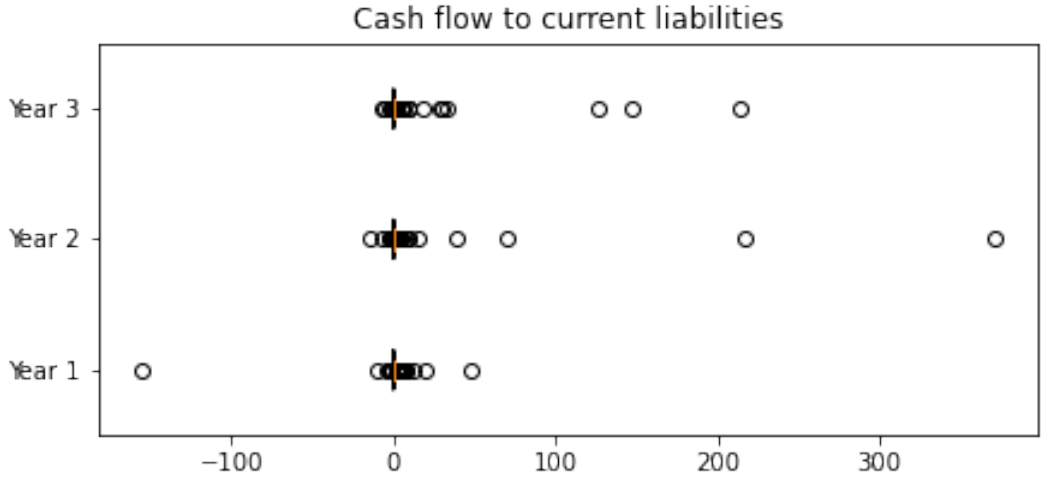
Due to the number of years for which the financial ratios were computed, they represent most of the variables included in the model. The distribution for each year of

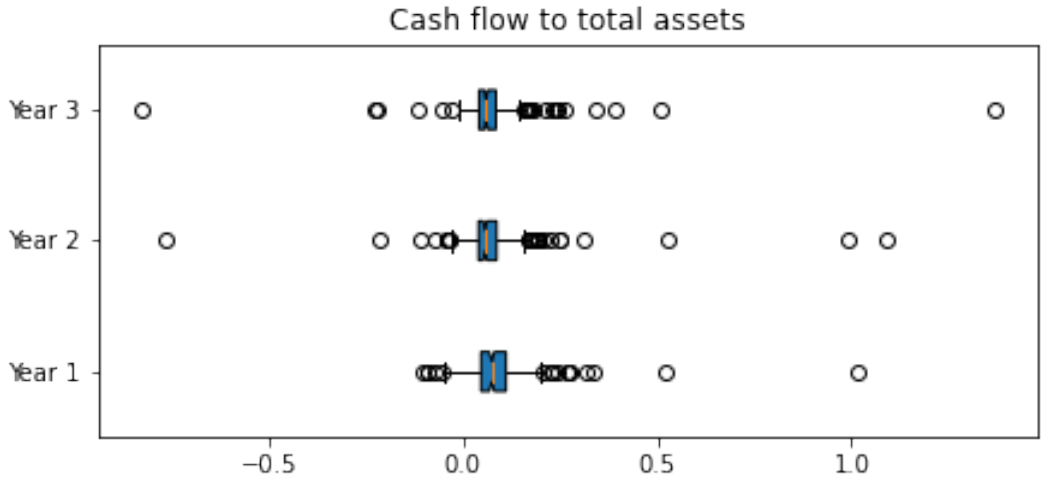
the ratios used is on the following charts:

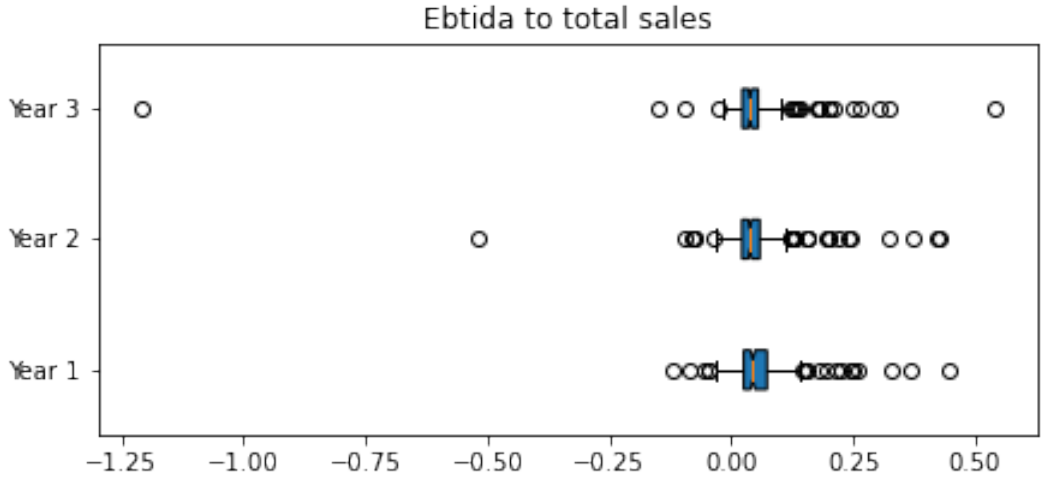



Ebitda to interest
Year 3
Year 2
Year 1
−60
−40
−20
0
20
40
60
80


Net current liabilites to cash flow
Year 3
Year 2
Year 1
0
20
40
60


Cash flow to current liabilities
Year 3
Year 2
Year 1
−100
0
100
200
300


Cash flow to total assets
Year 3
Year 2
Year 1
−0.5
0.0
0.5
1.0


Ebtida to total sales
Year 3
Year 2
Year 1
−1.25
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50

3 Methods

The main objective of this research was to verify which, among the main machine learning methods, would obtain the best results if they were used to replace the fundamentalist model of credit risk assessment applied by a human analyst.

3.1 Models

Ir order to find the model that best suits, we selected some benchmark Machine Learning models to test. After that, the hyperparameters were selected and the results obtained were evaluated, as described below.

3.1.1 Linear Regression

The first model tested was OLS, a linear regression, as it is a simple and widely used method in statistical studies. However, as expected, since predictor variables are a mixture of discrete and continuous variables, the linear discriminant analysis function did not prove to be optimal.

In this way, other models were used in order to seek better accuracy according to the metrics used. The linear techniques included in our study model the dependent variable as a linear function of the independent variables, whereas the non-linear techniques fit a non-linear model to the dataset.

3.1.2 Ridge, Lasso and Elastic Net

A linear model used was Ridge Regression, which solves a regression model imposing a penalty on the size of the coefficients (HOERL; KENNARD, 2000). It is a generic version of OLS with a shrinkage coefficient, which can be described as a parameter αthat controls the amount of shrinkage:

$$\min_{w} \|Xw - y\|_2^2 + \alpha\|w\|_2^2 \tag{3.1}$$

Ridge Regression can make the coefficients become more robust to collinearity, but it is not able to eliminate irrelevant variables.

We also used Lasso, which resembles the Ridge Regression, but instead of solving a $\ell 2$-penalized regression, it is a regression with a $\ell 1$-norm penalty. Basically, it consists on a linear model with an added regularization term:

$$\min_{w} \frac{1}{2n_{samples}} \|Xw - y\|_2^2 + \alpha \|w\|_1 \tag{3.2}$$

As other linear models, it minimizes the sum of squares, but it does variable selection and shrinkage (TIBSHIRANI, 2011). Lasso estimates sparse coefficients and is often used in some contexts when it is useful to reduce the number of features upon which the given solution is dependent.

Another linear model used here was Elastic Net, a model trained with both $\ell 1$ and $\ell 2$-norm regularization of the coefficients, combining the grouping effect of Ridge regression with the Lasso (ZOU; HASTIE, 2005). So we have:

$$\min_{w} \frac{1}{2n_{samples}} \|Xw - y\|_2^2 + \alpha\rho \|w\|_1 + \frac{\alpha(1-\rho)}{2} \|w\|_2^2 \tag{3.3}$$

Elastic-Net allows for learning a sparse model where few of the weights are non-zero like Lasso, while still maintaining the regularization properties of Ridge.

3.1.3 Ensemble models

Ensemble models were also used. One of them was Random Forest. It consists on randomized tree predictors where each tree in the ensemble is built from a sample drawn with replacement from the training set. It fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

As explained by Breiman 2001, a Random Forest consists of "a collection of tree-structured classifiers $\{h, (x, \Theta_k), k = 1, ...\}$, where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x". However, since we are working with a regression problem, instead of voting for the most popular class, the model computes the average of the forecasts.

Another ensemble model used was Gradient Boosting. It builds an additive model in a forward stage-wise fashion, allowing the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function. Its prediction $\hat{y}_i$, for a given input x_i, has the h_m *weak learners*. So its form is as follows:

$$\hat{y}_i = F_M(x_i) = \sum_{m=1}^{M} h_m(x_i) \tag{3.4}$$

According to Friedman 2001, one of the most most favorable aspect of this kind of procedure is robustness, since these procedures "are invariant under all (strictly) monotone

transformations of the individual input variables... as a consequence of this invariance, sensitivity to long-tailed distributions and outliers is also eliminated."

3.1.4 Other models

As stated before, since we are working with a small volume data, two other models were tested as well: kNN and a SVM regressor. kNN Regression is based on k-nearest neighbors, where the target is predicted by local interpolation of the targets associated of the nearest neighbors in the training set. Its decision surfaces are nonlinear, there is only a single integer parameter, which can be tuned with simple cross-validation. Its high capacity comes from the fact that it accesses the entire reservoir of training data at test time, but rarely causes overfitting itself (GOLDBERGER et al., 2004).

The basic idea of a Support Vector Machine is to "find a function $f(x)$ that has at most ε deviation from the actually obtained targets y_i for all the training data, and at the same time is as flat as possible" (SMOLA, 2004). In this manner, when used as a regressor, the model produced depends only on a subset of the training data, because the cost function ignores samples whose prediction is close to their target.

A Support Vector Regression solves the following primal problem:

$$\min_{w,b,\zeta,\zeta^*} \frac{1}{2} w^T w + C \sum_{i=1}^{n} (\zeta_i + \zeta_i^*) \tag{3.5}$$

$$\begin{aligned} \text{subject to} \quad & y_i - w^T \phi(x_i) - b \leq \varepsilon + \zeta_i, \\ & w^T \phi(x_i) + b - y_i \leq \varepsilon + \zeta_i^*, \\ & \zeta_i, \zeta_i^* \leq 0, i = 1, \ldots, n \end{aligned}$$

In SVM, different Kernel functions can be specified. In this research, we selected the polynomial function through cross-validation.

3.1.5 Hyperparameters selection

Since the hyperparameters are not directly learnt within estimators, to each model (that requires any hyperparameters) tested we used a grid search to select them. Grid search picks out a grid of hyperparameter values and evaluates all of them. In this process, we used 10-fold cross-validation to verify which hyperparameters would result in the best prediction. This way, constructed estimators were provided with optimized parameters.

3.2 Performance Metrics

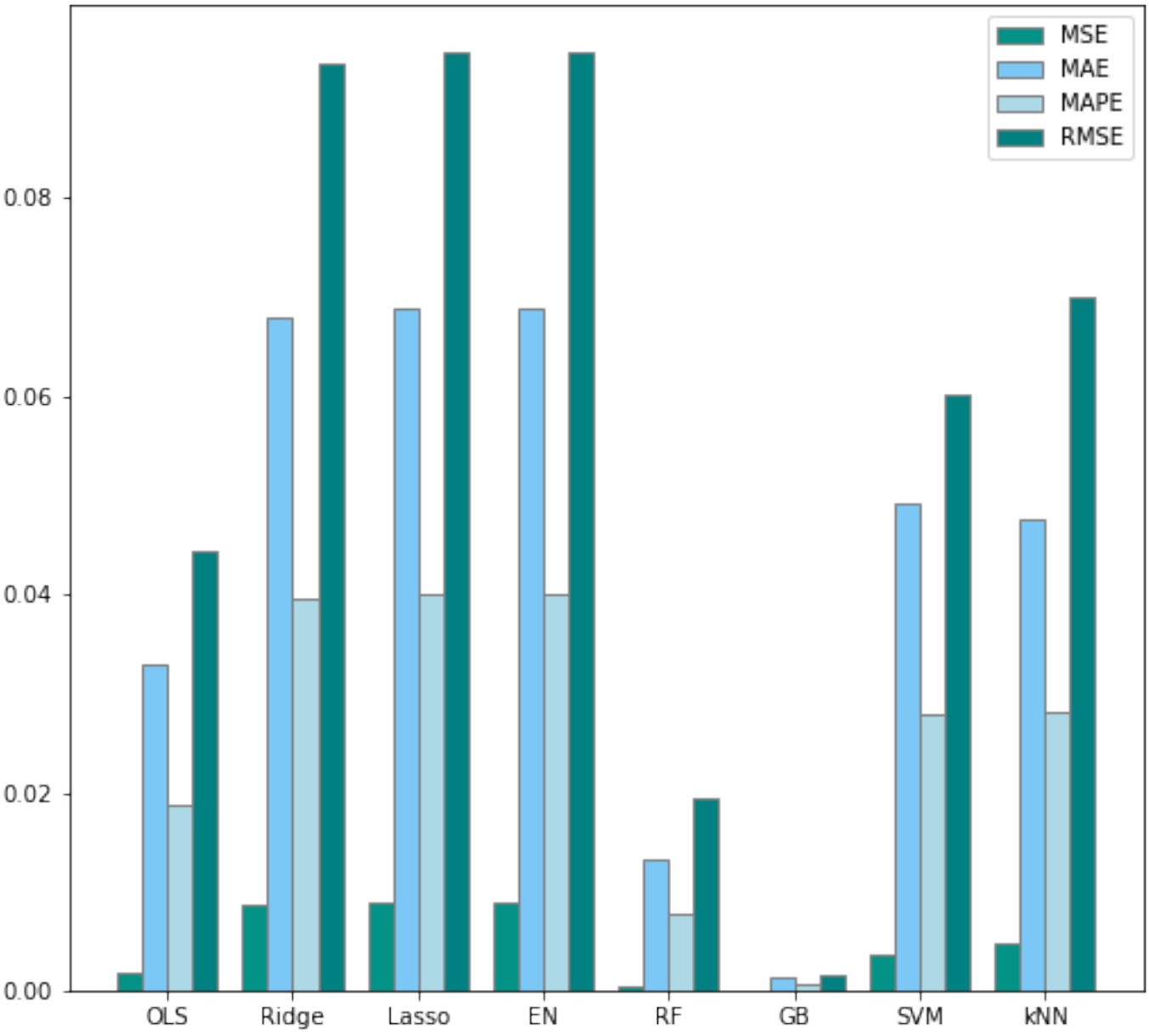
There are different performance metrics that can be used to evaluate the extent to which the models are more accurate on predicting the dependent variable. We chose four error metrics that are commonly used for evaluating and reporting the performance of a regression model, as listed on Table 1. Each one of the metrics listed has its own method of quantifying the model performance, and this is why we decided to keep all of them in the results presentation.

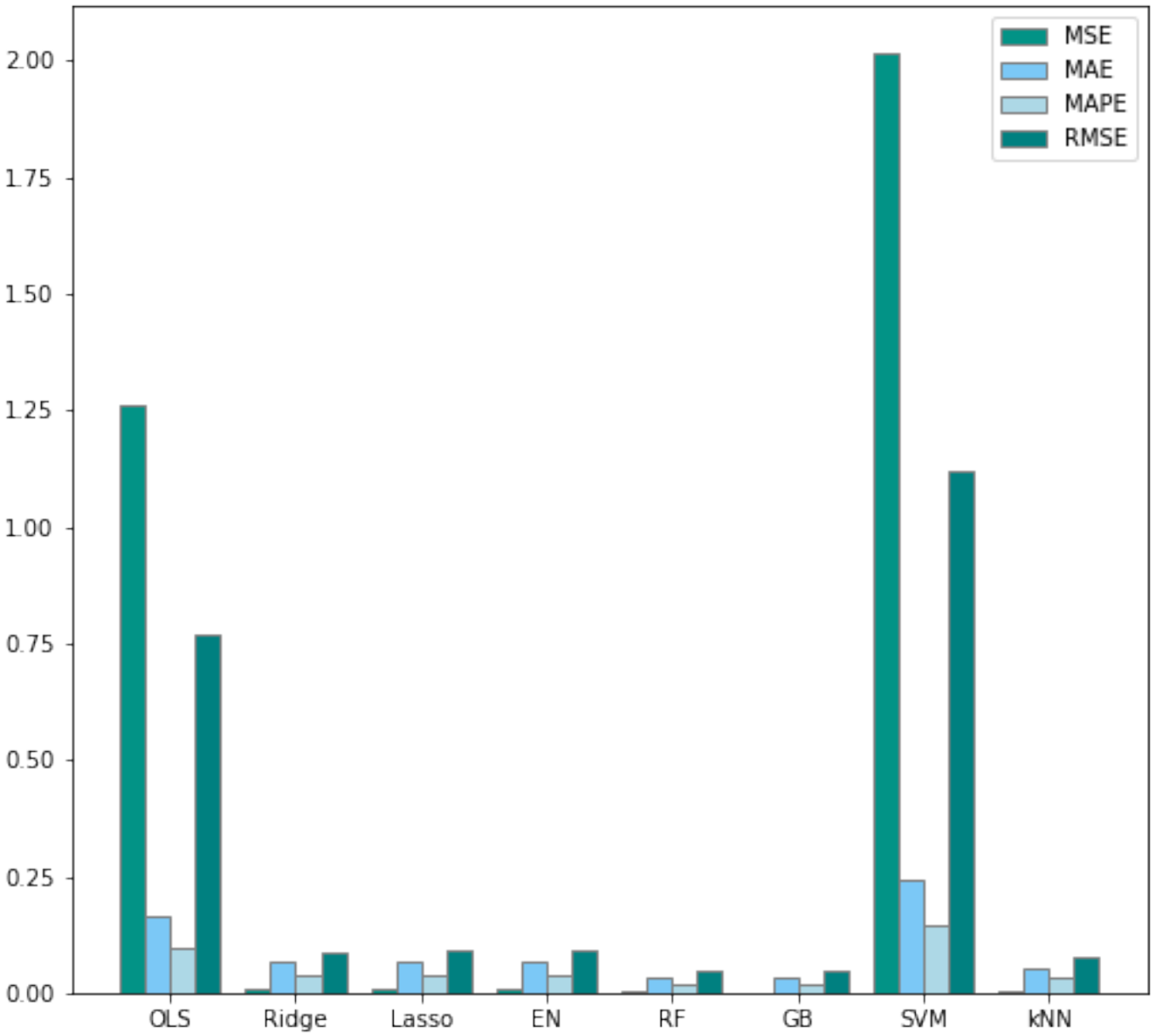
Metric
Mean absolute error regression loss (MAE)
Mean squared error regression loss (MSE)
Mean absolute percentage error regression loss (MAPE)
Root mean squared error regression loss (RMSE)

Table 1 – Metrics for Model Valuation

4 Results

We evaluated the results according to the metrics mentioned above and compared them for all the eight models tested. In the two following charts, we show the results for the train and test set.





The images above make it possible to compare the mean of all the four error metrics and look for the ones with the least error, since the smaller the error, the better the model. In this sense, it is clear that two of them had the best results: Gradient Boosting and Random Forests. However, Gradient Boosting still got a error's mean smaller for the test set. Through this comparison we can select it as the model that came up with the best outcomes.

In addition, the two models mentioned previously also have smaller deviations from the mean, including on the test set, as we can see on the table below:

Model	MSE		MAE	
	Train	Test	Train	Test
OLS	0.0003	2.0227	0.0023	0.1259
Ridge	0.0009	0.0033	0.0038	0.0132
Lasso	0.0009	0.0037	0.0040	0.0139
Elastic Net	0.0009	0.0038	0.0040	0.0140
Random Forest	2.75×10^{-5}	0.0009	0.0005	0.0052
Gradient Boosting	5.82×10^{-7}	0.0010	0.0001	0.0063
SVM	0.0003	2.6723	0.1677	0.0991
kNN	0.0004	0.0028	0.0025	0.0114

Model	MAPE		RMSE	
	Train	Test	Train	Test
OLS	0.0013	0.0756	0.0031	0.8186
Ridge	0.0024	0.0085	0.0047	0.0166
Lasso	0.0024	0.0090	0.0050	0.0.0187
Elastic Net	0.0025	0.0090	0.0050	0.0187
Random Forest	0.0003	0.0035	0.0007	0.0050
Gradient Boosting	7.86×10^{-5}	0.0041	0.0018	0.0097
SVM	0.0013	0.0991	0.0024	0.8744
kNN	0.0015	0.0075	0.0032	0.0169

Table 2 – Erro's Standard Deviation

Thus, regardless of the metric considered, the results for the test set show that Gradient Boosting proved to be the model with the best predictive capacity for this case, although Random Forest also obtained similar results and proved to be effective.

Each model also presented variations in the final amount of the most relevant variables, as well as in which variables fell into this category. In the case of the selected model, Gradient Boosting, we found that, from the 39 explanatory variables that we used as input for the model, only 6 variables had parameters lower than 0.001.

As expected, the most relevant variables are financial ratios, which are part of the C of Capital, such as leverage, liquidity and profitability ratios, especially from the last two years. Some registration aspects, which are part of the C of Character, such as the average amount kept in the bank account and complaints raised by suppliers against the cooperative, were also among the most relevant variables.

The following tables show the relevance of each variable calculated by permutation feature importance for the two models with best prediction results . The number of times a feature was randomly shuffled and returned a sample of feature importances was set to 30.

Feature	Mean	Std. Dev.
Net current liabilites to cash flow - Year 1	0.107	0.009
Results volatility	0.062	0.005
Current ratio - Year 1	0.033	0.003
Ebitda to interest - Year 3	0.022	0.002
Ebtida to total sales - Year 1	0.021	0.002
Cash flow to current liabilities - Year 2	0.021	0.002
Current plus long term liabilities to net worth - Year 1	0.021	0.002
Current ratio - Year 3	0.020	0.002
Overdue bank debts to total sales	0.019	0.002
Current bank debts to net worth - Year 1	0.015	0.003
Time operating - in months	0.014	0.001
Current plus long term liabilities to net worth - Year 2	0.013	0.001
Days since last restriction opened on cadastral surveys	0.012	0.001
Average value on the bank account	0.011	0.001
Cash flow to current liabilities - Year 3	0.011	0.001
Cooperative strategy not clear	0.011	0.002
Ebtida to total sales - Year 2	0.011	0.001
Ebitda to interest - Year 1	0.008	0.001
Have a system to grant credit to cooperative's members	0.008	0.002
Current bank debts to net worth - Year 2	0.007	0.000

Table 3 – Permutation Features Importance - Gradient Boosting

Feature	Mean	Std. Dev.
Net current liabilites to cash flow - Year 1	0.063	0.005
Cash flow to total assets - Year 1	0.041	0.004
Ebitda to interest - Year 1	0.038	0.003
Results volatility	0.034	0.004
Current ratio - Year 1	0.030	0.002
Ebitda to interest - Year 2	0.029	0.003
Current plus long term liabilities to net worth - Year 2	0.029	0.002
Current plus long term liabilities to net worth - Year 1	0.028	0.003
Current ratio - Year 2	0.027	0.002
Ebtida to total sales - Year 1	0.026	0.003
Ebitda to interest - Year 3	0.021	0.002
Current bank debts to net worth - Year 1	0.019	0.002
Current ratio - Year 3	0.019	0.002
Cash flow to total assets - Year 2	0.017	0.002
Ebtida to total sales - Year 2	0.017	0.001
Net current liabilites to cash flow - Year 2	0.016	0.002
Current bank debts to net worth - Year 2	0.013	0.001
Current plus long term liabilities to net worth - Year 3	0.011	0.001
Overdue bank debts to total sales	0.010	0.001
Cash flow to current liabilities - Year 1	0.009	0.001

Table 4 – Permutation Features Importance - Random Forest

As we can see on the tables above, Gradient Boosting was more concentrated on less variables than Random Forest. For Gradient Boosting, four features proved to be without any relevance, while for the second model only three variables had a score equal to zero. This way, it is possible to concentrate more on features that really affect the result in future assessments.

5 Conclusion

Our work uses well known Machine Learning models in substitution of fundamentalists assessments made by analysts to measure the credit risk of farmers cooperatives. We used data easily obtained from different sources and tried to predict the final score of the cooperatives as the explained variable. We tested eight models: Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Random Forest, Gradient Boosting, SVM and kNN.

We had great results using different models, and more than one could be used to predict farmers cooperative credit score using the input items that we had. However, among the models tested, Gradient Boosting achieved the best results, followed closely by Random Forest.

The outcome of this study shows that Machine Learning is evolving and tends to be increasingly used in credit risk problems. This may happen not only in classification problems, when the objective is to separate good and bad borrowers, but also in regression problems when we need to achieve a continuous variable: in this case, the cooperative's score.

As we did in this case, in which we gathered borrowers with similar nature and characteristics, more studies could be conducted for other types of companies, taking into consideration their specificities, instead of comparing a big number of samples with very distinct situations. The objective of this method is to choose the model that best fits the set of companies evaluated.

Since there are many new models that can also be tested for this case, in which we have a sample with a reduced number of observations, further studies can improve the use of Machine Learning techniques for the problem. The models used here themselves may have wider hyperparameter tests, which could not be performed due to computational restrictions.

Bibliography

ADDO, P. M.; GUEGAN, D.; HASSANI, B. Credit risk analysis using machine and deep learning models. *Risks*, v. 6, n. 2, 2018. ISSN 2227-9091. Disponível em: <https://www.mdpi.com/2227-9091/6/2/38>. Quoted on page 15.

ANICETO, M. C.; BARBOZA, F.; KIMURA, H. Machine learning predictivity applied to consumer creditworthiness. *Future Business Journal*, v. 6, 2020. ISSN 2314-7210. Disponível em: <https://doi.org/10.1186/s43093-020-00041-w>. Quoted on page 14.

BEAVER, W. H. Financial ratios as predictors of failure. *Journal of Accounting Research*, [Accounting Research Center, Booth School of Business, University of Chicago, Wiley], v. 4, p. 71–111, 1966. ISSN 00218456, 1475679X. Disponível em: <http://www.jstor.org/stable/2490171>. Quoted on page 16.

BREEDEN, J. L. Survey of machine learning in credit risk. 2020. Disponível em: <http://dx.doi.org/10.2139/ssrn.3616342>. Quoted on page 15.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, p. 5–32, 2001. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1010933404324>. Quoted on page 23.

BUSSMANN, N. et al. Explainable machine learning in credit risk management. *Computational Economics*, v. 57, p. 203–216, 2021. ISSN 1572-9974. Disponível em: <https://doi.org/10.1007/s10614-020-10042-0>. Quoted on page 15.

CHOW, J. *Analysis of Financial Credit Risk Using Machine Learning*. Tese (Doutorado), 04 2017. Quoted on page 16.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 29, n. 5, p. 1189–1232, 2001. ISSN 00905364. Disponível em: <http://www.jstor.org/stable/2699986>. Quoted on page 23.

GOLDBERGER, J. et al. Neighbourhood components analysis. MIT Press, v. 17, 2004. Disponível em: <https://proceedings.neurips.cc/paper/2004/file/42fe880812925e520249e808937738d2-Paper.pdf>. Quoted on page 24.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, [Taylor Francis, Ltd., American Statistical Association, American Society for Quality], v. 42, n. 1, p. 80–86, 2000. ISSN 00401706. Disponível em: <http://www.jstor.org/stable/1271436>. Quoted on page 22.

KHANDANI, A. E.; KIM, A. J.; LO, A. W. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking Finance*, v. 34, n. 11, p. 2767–2787, 2010. ISSN 0378-4266. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0378426610002372>. Quoted on page 14.

LEO, M.; SHARMA, S.; MADDULETY, K. Machine learning in banking risk management: A literature review. *Risks*, v. 7, n. 1, 2019. ISSN 2227-9091. Disponível em: <https://www.mdpi.com/2227-9091/7/1/29>. Quoted on page 15.

SHOUMO, S. Z. H. et al. Application of machine learning in credit risk assessment: A prelude to smart banking. In: *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*. [S.l.: s.n.], 2019. p. 2023–2028. Quoted on page 15.

SMOLA, B. S. A. J. A tutorial on support vector regression. *Statistics and Computing*, v. 14, n. 14, p. 199–222, 2004. ISSN 1573-1375. Disponível em: <https://doi.org/10.1023/B:STCO.0000035301.49549.88>. Quoted on page 24.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, [Royal Statistical Society, Wiley], v. 73, n. 3, p. 273–282, 2011. ISSN 13697412, 14679868. Disponível em: <http://www.jstor.org/stable/41262671>. Quoted on page 23.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, [Royal Statistical Society, Wiley], v. 67, n. 2, p. 301–320, 2005. ISSN 13697412, 14679868. Disponível em: <http://www.jstor.org/stable/3647580>. Quoted on page 23.

Annex

ANNEX A – Features

Feature
Overdue bank debts to total sales
Days since last restriction opened on cadastral surveys
Other outstanding debts to total sales
Amount of restrictions on cadastral surveys
Days since last restriction on board member's name opened on cadastral surveys
Days since bank account has been opened
Average value on the bank account
Time operating - in months
Results volatility
Current plus long term liabilities to net worth - Year 1
Current plus long term liabilities to net worth - Year 2
Current plus long term liabilities to net worth - Year 3
Current bank debts to net worth - Year 1
Current bank debts to net worth - Year 2
Current bank debts to net worth - Year 3
Net current liabilites to cash flow - Year 1
Net current liabilites to cash flow - Year 2
Net current liabilites to cash flow - Year 3
Current ratio - Year 1
Current ratio - Year 2
Current ratio - Year 3
Ebitda to interest - Year 1
Ebitda to interest - Year 2
Ebitda to interest - Year 3
Cash flow to current liabilities - Year 1
Cash flow to current liabilities - Year 2
Cash flow to current liabilities - Year 3
Cash flow to total assets - Year 1
Cash flow to total assets - Year 2
Cash flow to total assets - Year 3
Ebtida to total sales - Year 1
Ebtida to total sales - Year 2
Ebtida to total sales - Year 3
Information lacking or poor quality
Have a system to grant credit to cooperative's members
Cooperative strategy not clear
Hedge against price's variation
Products diversity
High Risk on the sector