# Credit classification using regulation techniques on the Credit German database

*Sebastián Sosa*[1], *Priscila Rivera*[1], *Cristhian Aldana*[1], *Yesenia Saavedra*[1], *Luis Trelles*[1], *Gustavo Mendoza*[1]

[1] Universidad Nacional de Frontera - UNF
Av. San Hilarión 101, Sullana, Piura, PERÚ

## ABSTRACT

The development of microfinance, as well as microcredit, has generated greater competition among financial institutions to attract customers in this business segment. For this reason, the development of credit scoring models is highly required by these financial institutions. In this sense, to ensure that no overfitting is generated in the use of prediction techniques and in case of difficulty with collinearity, it will not be possible to obtain reliable estimates and predictions through common statistical techniques such as least squares; for this reason, it is significant and necessary to apply regularized regression methods such as Ridge, Lasso and Elastic Net. The present research determined the optimal credit scoring model for the Credit German database using the Ridge, Lasso, and Elastic Net regulation techniques. This dataset was initially analyzed with the Logit model, finding that this model has an accuracy of 37.2%; on the other hand, the Lasso model presented an accuracy of 76.7%, the Ridge model of 75.6%, and the Elastic Net model of 69.2%. Finally, the Lasso model evidenced the best prediction of the credit rating of Credit German future clients, with an accuracy in the training data of 82.9% and for the test data of 76.7%, being superior to the proposed models.

**Keywords**: Credit scoring, Ridge, Lasso, Elastic Net, Beta, Logit model

# INTRODUCTION

The regulation techniques such as Ridge, Lasso, and Elastic Net, allow avoiding the negative impacts that a collinearity problem would generate in a linear model estimated by the Least Squares technique. In this sense, the Ridge method contracts the regression coefficients when the penalty term is included in the objective function. On the other hand, the Lasso method is a regularized linear regression technique, like Ridge, with a slight difference in the penalty zero estimates for some coefficients and non-zero for others, whereby Lasso performs a kind of continuous variable selection, due to the L1 rule used in its formulation. Lasso reduces the variability of the estimates by reducing the coefficients and at the same time produces interpretable models by reducing some coefficients to zero.

It should be noted that Lasso has a greater advantage over Ridge regression in variable selection because it produces simpler and more interpretable models involving a single subset of the predictors. However, no one method always dominates the other. Generally, one might expect Lasso to be better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or equal to zero. On the other hand, the Ridge technique might perform better when the response is a function of many predictors, all with coefficients of approximately the same size. However, the number of predictors that are related to the response is not known a priori for real data sets. A technique such as cross-validation can be used to determine which approach is best in a particular data set. Like Ridge regression, when least-squares estimates have excessively high variance, the Lasso solution can produce a reduction in variance at the expense of a small increase in bias, and thus can generate more accurate predictions.

A financial company is an entity that has as its main objective to receive the greatest amount of capital resources in order to have solvency and facilitate both services and goods, through credit operations (Tigreros, 2020). Financial institutions consider important the liquidity that allows them to fulfill their obligations to lend credit to their customers. These institutions also aim to avoid the lower risk they face when approving a loan. In Peru, the good development and boom of Microfinance have been based on important points such as transparency of risk centers and financial reporting standards, appropriate credit technology systems, a regulatory framework of institutional support, high promotion of price transparency, and final competition in the market. However, these institutions are threatened by the entry of new oligopolistic companies with greater capital and, of course, better technologies for the detection of credit risks (Quiroz, 2020).

That is why it is essential to develop mechanisms that help reduce this problem of credit risk, in order to identify customers who do not comply with their obligations, that is, who present insolvency to satisfactorily complete a loan. Well, the biggest concern of these financial institutions is to end up in legal cases in order to recover

only their capital and devise portfolio management processes for debtor clients (Quin et al. 2021). This paper analyzes the dataset of the German private loan entity called Credit German, applying regulatory techniques (Lasso, Ridge, and Elastic Net) in order to determine the credit rating forecast for future credit customers of Credit German, to achieve this we will identify the best model with the regulation techniques for credit rating, we will analyze the credit attributes of Credit German customers.

When evaluating the credit risk of a financial institution, for example in a cooperative, the Logit model can be used, which works with a binary type variable, so that when the number tends to 1 it can be said that the associated clients have a good performance, and on the contrary 0 represents that they do not have a good performance in their portfolio. Likewise, the risk factors for male members may be their young age or their single marital status, as well as their low resources, or the limited time they have belonged to the cooperative, working in private entities, and having large loans with long repayment terms. In a case study, it was found that the percentage of good hits of the model identified was 63.95%, i.e., the level of prediction is acceptable, and it was found that members reduce the risk of default by 0.45% each year and if they increase their income, for each additional 1% their level of default decreases by 6.74%. Finally, the Logit model allows as a quantitative tool to predict possible risks (Pardo and Díaz, 2020).

By analyzing the dataset of a financial institution, it is possible to classify its customers, based on regression models, methodologically applying a support vector machine that allows to classify users as "bad or good" and thus make the credit assigned to each one of them (Guevara, 2020). On the other hand, the credit risk analysis of a Credit Scoring model for a company can use the Delphi study method, which allows establishing both quantitative and qualitative variables to resort to the evaluation of the quality of each client, being the dimensions of some of its qualitative variables: legal and commercial issues, seniority and contribution of the company, among others (Leal et al. 2018). There are different techniques or methods to carry out credit rating, which generate greater efficiency; then, from subsets, it is possible to adjust such data for the improvement of the prediction performance. Here, theoretically one can use the classifiers corresponding to the gradient increase at the extreme level using the random forest (Honglian, 2018). For “Cajas Municipales”, credit scoring models reduce credit risk, using binary logistic regression considering a dichotomous discrete dependent variable. In addition, the model allows to statistically qualify and make predictions with a high percentage of accuracy of the credits granted.

In emerging countries such as Peru, different small and medium-sized companies make up a very important productive sector in the competitive development of the country; however, their performance and evolution are not sustainable over time, since many do not have adequate support that allows them to strengthen their development and growth; since many times the limitations of access to different formal financing sources make it impossible for them to manage an adequate development and administration of operations. Sometimes, these difficulties may be

due to a lack of timely and didactic financial information, and a lack of an adequate methodology that considers the multiple realistic variables involved in this type of model (Siguas, 2018). Thus, the purpose of the credit scoring model is to evaluate credit risk in financial institutions, since credit evaluation can often be slow and deficient (Flores, 2019).

Finally, the present work evaluated a credit scoring model for the Credit German database, using the regulation techniques: Ridge, Lasso, and Elastic Net. Likewise, Logit was used as the main model, analyzing its, parameters based on it, so that from there, in future works, the credit rating forecast for future Credit German clients can be determined. Among the variables used we have the credit rating of Credit German's future customers (with a dichotomous scale of 1: Good credit performance and 0: Bad credit performance or default for the Logit model) based on the following attributes or independent variables such as collateral, period, age, seniority, bachelorhood and normal rating, which in the delinquency of the given credits are significantly statistical; on the other hand, it is also possible to work with non-representative variables such as woman, amount, hectare, among others. It should be noted that some variables such as period, singleness, age, age guarantee, seniority, singleness, and normal qualification, can serve as socioeconomic determinants (Romero, 2017).

# MATERIALS AND METHODS

As a methodology, regulation models were applied as a technique to apply a model whose response variable is of a binary type where with values of 0 when the client is not suitable for credit and 1 when he can access credit. The population is 1000 customers of the Credit German database that is located within the UCI machine learning repository. The R programming language and the RStudio code editor were used for data processing and precision, ROC curve and accuracy were used. The model for Lasso regression is as follows:

$$\hat{\beta}LASSO = argmin\left\{\sum_{i=1}^{n}\left(yi - \beta_0 - \sum_{j=1}^{\mathrm{p}}\beta_i x_{il}\right)^z + \lambda\sum_{j=1}^{p}|\beta_i|\right\} \qquad (1)$$

In which the value of causes the parameters of the model to take the value of 0. For the Ridge model the following model is expressed:

$$\hat{\beta}RIGDE = argmin\left\{\sum_{i=1}^{n}\left(yi - \beta_0 - \sum_{j=1}^{\mathrm{p}}\beta ixij\right)^z + \lambda\sum_{j=1}^{p}\beta_i^2\right\} \qquad (2)$$

In which the value of makes the model parameters approximate 0. The Elastic-Net model is represented as follows:

$$\frac{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2}{2n} + \lambda\alpha\sum_{j=i}^{p}|\beta_i| + \frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2 \qquad (3)$$

The Elastic-Net model is a combination of the Ridge model as the Lasso model, the parameter values are within the range of 0 to 1. When the parameter α takes the value of zero it is a Ridge regression and when it is equal to 1 it is a Lasso regression. Of the 20 attributes that can be identified in the Credit German database (current account status, duration in months, credit history, purpose, amount of credit, savings/bond account, current employment since (time), payment rate as a percentage of disposable income, personal status and gender, other debtors/guarantors, current residential, property, age in years, other installment payment plans, housing, number of existing loans at this bank, job, number of persons eligible for maintenance, telephone, and foreign worker) were selected (Kennedy, 2013).

# RESULTS

Of the 20 attributes that can be identified in the Credit German database (Current account status, duration in months, credit history, purpose, amount of credit, savings/bond account, current employment since (time), payment rate as a percentage of disposable income, personal status, and gender, other debtors/guarantors, current residential, property, age in years, other installment payment plans, housing, number of existing credits in this bank, work, number of people susceptible to maintenance, telephone, and foreign worker) 3 of them were selected as a sample for the respective processing among them the credit history.

To ensure the integrity of the dataset (degree of missing values) in the present investigation, outliers were processed using the multivariate form of variance of the k-prox method. At the end of it, from the data of 1000 customers, 25 outliers remained, which were simplified because they biased the developed analysis. On the other hand, the data has a class imbalance, class 1 represented by those people who are approved for the loan exceeds class 0 which represents people who are not accessible to the loan. This type of problem generates a bias in the model since it is trained with a larger amount of one class, this makes it unable to detect with better accuracy minority class, to solve this, balancing techniques are applied to the training data (Parker et al., 2006).

After having carried out the preprocessing of the data, the models begin to be applied, the first model is logistic regression (Logit), this model is used because it is about classifying those people who can be qualified to a credit. Now they begin to apply the regulation models for the selection of predictors in the case of the regression Ridge only approaches the predictors β to 0, the regression Lasso tries to select parameters giving as value to the β and the Elastic-Net regression model combines the methods of the two techniques in order to overcome its limitations (Santana et al., 2017).

The values of each parameter of the models already mentioned are observed; the logit model is taken because it is also taken within the linear regression models and is also used for this type of problem consisting of classification (see Table 1).

Table 1. Parameter values for each model

| Model | λ | A |
|---|---|---|
| Logit | - | - |
| Lasso | 0,0178148 | 1 |
| Ridge | 0,1286373 | 0 |
| Elastic-Net | 0,128461 | 0,1 |

After pulling the parameters of the model, its performance begins to be evaluated for both the training data and the test data (see Table 2).

Table 2. Train Validation

| Modelo | Precisión | Accuracy |
|---|---|---|
| Logit | 0.811 | 0.834 |
| Lasso | 0.829 | 0.806 |
| Rigde | 0.829 | 0.806 |
| Elastic-Net | 0.792 | 0.813 |

We can see in this table that the model that performs in a better way are 2 which are the Lasso model and the Ridge model, this is with the training data. It is now evaluated with the test data (see Table 3).

Table 3. Validation test

| Model | Precision | Accuracy |
|---|---|---|
| Logit | 0.372 | 0.265 |
| Lasso | 0.767 | 0.714 |
| Rigde | 0.756 | 0.704 |
| Elastic-Net | 0.692 | 0.711 |

In this case the best model is the Lasso model because it has a higher accuracy than the other models with a value of 0.767. This can also be assessed with the ROC curve presented (see Figure 1).

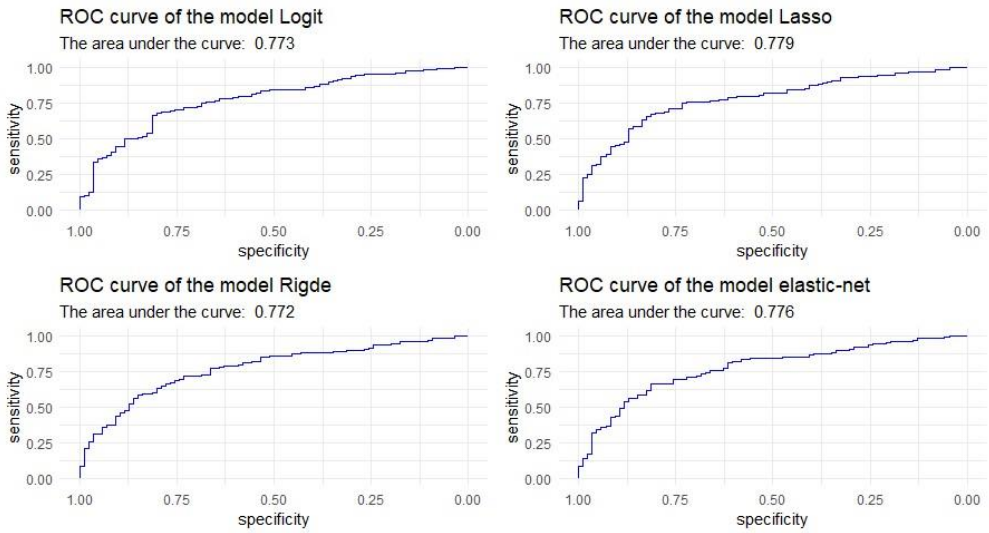


Figure 1. ROC-Models Curve The model with the highest AUC value is the Lasso model. (A, Logit; B, Lasso; C, Ridge; D, Elastic-Net)

# CONCLUSIONS

The most widely implemented credit scoring model is the Logit model; however, to ensure that no overfitting is generated in the use of prediction techniques and in case of difficulty with collinearity, it will not be possible to obtain reliable estimates and predictions through common statistical techniques such as least squares or model logit; for this reason, it is significant and necessary to apply regularized regression methods such as Ridge, Lasso and Elastic Net (Carrasco 2016). This paper reveals that Model Logit is acceptable since its percentage of successes is 63.95%, on the other hand, the studies that used regression techniques to perform the risk of credit analysis and other techniques such as machine learning (López and Maldonado, 2019). The most accurate model in this research is the Lasso model, with an accuracy of 76.7%, this being a parametric method. Other authors use non-parametric models that use a model of vector support machines that has an accuracy of 85.1%. The overall conclusion obtained in this work is that the techniques used for the study and modeling of non-compliance for a client that belongs to this database, the above, are supported by a high index of accuracy. Additionally, it is considered that, within these proven techniques, the best in the regulation technique is Lasso which has a high accuracy of 0.7667 and the area under the curve is equal to 0.779 which tells us that it predicts in a better way. These models exceed the logit model with high pressure, because according to the test data the Logit model has an accuracy of 0.372, compared to the Lasso and Ridge models that present 0.767 and 0.756 in precision respectively. Finally, it is worth mentioning that these models use 20 variables of which only three were selected.

## ACKNOWLEDGMENTS

The authors would like to acknowledge the Universidad Nacional de Frontera, Sullana, Piura, Perú.

## REFERENCES

Carrasco, M. (2016). Regression regularization techniques: implementation and applications. Universidad de Sevilla. Mathematics

Flores Bardales, A. (2019). Methodology for implementing Credit Scoring in a financial institution in the SME segment.

Guevara, K. (2020). Credit classification using vector support machines on the Lending Club database. OBSERVATORY OF ECONOMICS AND NUMERICAL OPERATIONS Volume 59.

Honglian, H. (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios, ELSEVIER Volume 98.

Kennedy, K. (2013). Credit Scoring Using Machine Learning. Technological University Dublin. Science.

Leal, A., Aranguiz, M. y Gallegos, J. (2018). Credit risk analysis, due to the Credit Scoring model. JOURNAL FACULTY OF ECONOMIC SCIENCES Volume 26 No. 1. pp. 181-207.

López, J. and Maldonado, S. (2019) 'Profit-based credit scoring based on robust optimization and feature selection', Information Sciences, 500, pp. 190–202.

Pardo Carrillo, O. S., & Díaz Castro, J. (2020). Credit risk profile for a cooperative in Villavicencio based on a Logit model. Revista Universidad y Empresa, Volume 22 No. 38.

Parker, M., Moleshe, V., De la Harpe, R. & Wills, G. (2006). An evaluation of information quality frameworks for the world wide web. In Proceedings of the 8th Annual Conference of WWW Applications. 70

Qin, C., Zhang, Y., Bao, F., Zhang, C., Liu, P., & Liu, P. (2021). XGBoost Optimized by Adaptive Particle Swarm Optimization for Credit Scoring. Mathematical Problems in Engineering.

Quiroz, M. (2020). Credit scoring a tool to minimize the credit risk of microfinance institutions-Peru. QUIPUKAMAYOC Volume 69 No. 76.

Romero, L.C. (2017). The credit scoring model as an alternative for credit evaluation in Agrobanco. QUIPUKAMAYOC Volume 25 No. 49. pp. 101-109.

Santana, P.., Villa Monte, A., Rucci, E., Lanzarini, L. y Fernández, A. (2017), Analysis of Methods for Generating Classification Rules Applicable to Credit Risk. Journal of Computer Science and Technology, Vol. 17, núm.01, pp.20-28

Siguas, A. (2018). The process of granting credit and its relationship with the credit classification of the debtor in the company of goods and services for the home S.A.C., Perú.

Tigreros, D. (2020). Prediction of credit risk in Colombia using artificial intelligence techniques. UIS Engineering Journal, 37-52.