1-6-2021

# The Effects of Customer Segmentation, Borrowers' Behaviours and Analytical Methods on the Performance of Credit Scoring Models in the Agribusiness Sector

Daniela Lazo
*Universidad de Talca*, dlazo09@alumnos.utalca.cl

Raffaella Calabrese
*The University of Edinburgh*, Raffaella.Calabrese@ed.ac.uk

Cristian Bravo Roman
*Western University*, cbravoro@uwo.ca

# The effects of customer segmentation, borrowers' behaviours and analytical methods on the performance of credit scoring models in the agribusiness sector*

Daniela Lazo†, Raffaella Calabrese‡, Cristián Bravo§

**Abstract**

The main aim of this study is to analyse the joint effects of customer segmentation, borrowers' characteristics and modelling techniques on the classification accuracy of a scoring model for agribusinesses. To this end, we used data provided by a Chilean company on 161,163 loans from January 2007 to December 2013. We considered random forest, neural network and logistic regression models as analytical methods. Regarding the borrowers' profiles, we examined the effects of socio-demographic, repayment-behaviour, agribusiness-specific and credit-related variables. We also segmented the customers as individuals, SMEs and large holdings. As the segments show different risk behaviours, we obtained a better performance when we estimated a scoring model for each segment instead of using a segmentation variable. In terms of the value of each set of variables, behavioural variables increased the predictive capability of the model by double the amount achieved by including agribusiness-related variables. The random forest is the model with the best classification accuracy.

**Keywords:** Agribusiness finance, credit scoring, repayment behaviour, random forests, logistic regression

## 1 Introduction

The main focus of this paper is the credit risk assessment in the agricultural sector. We use the definition of agriculture provided by the International Standard Industrial Classification of All

*NOTICE: this is the author's version of a work accepted for publication. Changes resulting from the publishing process, such as editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. Please cite this paper as follows: Lazo, D., Calabrese, R., Bravo, C. (2020) The Impact of Borrower Behaviour, Sectorial Variables and Modelling Techniques on Credit Risk Assessment in the Agribusiness Sector. The Journal of Credit Risk 16 (4): 119-156. DOI: https://doi.org/10.21314/JCR.2020.272. 

†Corresponding author. Universidad de Talca, Curicó, Chile. dlazo09@alumnos.utalca.cl

‡Business School, University of Edinburgh, United Kingdom.

§Department of Statistical and Actuarial Sciences, The University of Western Ontario, Canada.

Economic Activities (ISIC); the definition includes crops and livestock production, forestry, and hunting and fishing (United Statistics Division, 2016). Agricultural production is an inherently risky business whose risks also affect the lenders providing financial leverage to the sector. The variability of farm outputs is mainly explained by production risk (Tiedemann & Latacz-Lohmann, 2013), which comprises not only financial risks but also damage by pests, diseases and weather effects that can result in the borrowers' inability to repay their loans (Hazell, 1992). Moreover, agriculture has long production cycles, during which the market prices of agricultural products may diverge rapidly from initial projections (Becerra, 2004). Finally, agricultural lending is subject to a relatively higher moral hazard risk, both because farmers have more knowledge about their production risks than their lenders do and because information about borrowers with low incomes, which is common in small farming, is difficult to obtain (Becerra, 2004).

Agribusinesses have a pressing need for funding in order to sustain their operations. Given that the production cycle can be of a year or more, the need for working capital and the funding to acquire the supplies needed to operate within the cycle is usually supported by loans. The providers of these loans must carefully control their credit risk, but detailed studies about how to deal with this risk are not common.

Most of the studies on credit risk for primary producers have used financial ratios such as liquidity, profitability and leverage (Jouault & Featherstone, 2011; L. H. Miller & LaDue, 1988; M. P. Novak, LaDue, et al., 1999; Rambaldi, Zapata, Christy, et al., 1992). Gallagher (2001) reported that the inclusion of non-financial agribusiness-related characteristics brought significant improvements to the model. Hou, Skees, and Wang (2005) included demographic statistics and loan information such as loan size and lending year, providing a higher number of significant variables. Limsombunchai, Gan, and Lee (2005) defined the lending decision as a function of borrowers' characteristics, relationship indicators and dummy variables about the agricultural sector and loan information. Aruppillai and Phillip (2014) showed that considering socio-economic characteristics, such as number of family members, amount of loan disbursement and secondary education, improves the efficiency of the lending decision.

In addition to the choice of variable sets to include, another important aspect of credit risk assessment in the agribusiness sector is segmentation, that is, dividing the clients into groups according to a specific variable or set of variables. In some cases, using several scorecards on different customer segments provides better risk differentiation than using just one scorecard on everyone (Siddiqi, 2007). In credit scoring for the agribusiness sector, there are various possible segmentations; for example, current and non-current loans (Ziari, Leatham, & Turvey, 1994), loan size (L. H. Miller & LaDue, 1988), type of activity or produce (Bandyopadhyay, 2008), or loan type (Bandyopadhyay, 2008).

We performed a credit risk analysis on a dataset provided by a Chilean company that grants credit to farmers. This company is a major distributor of agricultural supplies, machinery and services to support farmers (businesses and people) in Chile. The company business has an im-

portant seasonal component. In effect, income and costs are more concentrated in the second half of each year.

One of the most important services offered by this company is the granting of credit to pay for supplies. The company gives financial alternatives that fit the farmer's needs, considering, for example, the seasonality of crops. To manage the credit risk, the company has credit and collection policies that are controlled regularly, but it does not have any automatic model to support the credit risk process.

The company offers instalment loans in payment structures equivalent to agricultural cycles, and most of them have terms of less than one year (99%). Around 90% of the loans are insured; however, depending on the credit policies, additional guarantees, such as mortgages or personal guarantees, could be required.

Regarding credit risk research in Chile, there are a few studies about it. Romani et al. (2002) used different techniques to predict bankruptcy in Chilean companies and found that neural networks performed better than logistic regression and discriminant analysis. Fica, Casanova, and Mardones (2018) concluded that a credit scoring model allowed greater flexibility and objectivity in the credit management process in a company dedicated to the production, marketing and distribution of asphalt products in the southern zone of Chile. Madeira (2019) indicated that the default rate of the total consumer loan portfolio of all Chilean banks has a high covariance risk and recommends that banks reduce the default rate of their loan portfolio by choosing customers that suffer fewer shocks during economic downturns.

To the best of our knowledge, our study is the first to analyse the joint effects of modelling techniques, segmentation and borrowers' characteristics on the performance of scoring models in the agribusiness sector. Specifically, the information available on the borrowers is socio-demographic data and repayment behaviour, in addition to agribusiness-specific and credit-related variables. We use the data of a company that grants credit and distributes supplies. Funding sources that also serve as input suppliers (with multiple offices close to their customers) have the advantage of being geographically close to customers and having knowledge of different agricultural specialities (ODEPA, 2013).

Because the customers had different sizes of agricultural crops and varied incomes, we could segment them and compare the different types of clients. We also analysed the impacts of the available information on farmers, measuring the contribution of these variables to default prediction. Finally, we examined the main classification techniques used in the industry and in the literature (Thomas, Crook, & Edelman, 2017) to understand the value of using a complex, non-linear, technique such as a neural network or random forest, instead of a simpler technique such as a logistic regression.

Each of these factors was determinant to build an efficient model. In order of importance, better information was, unsurprisingly, the top factor that can be used to improve predictive models, followed by the choice of model (analytical technique) and the segmentation (size of the

company). One caveat related to the last factor is that holdings require their own model because they are structurally different from both large (non-holding) companies and small farmers.

The organisation of the document is as follows. In the following chapter, we review the literature on agricultural credit scoring and explain the main financing sources available for farmers. Second, we describe the data, and we present the main credit scoring methodologies. Third, we show the empirical results. Finally, we draw conclusions, state the limitations of the research and suggest directions for future work.

# 2 Measuring Credit Risk in the Agribusiness Sector

Credit risk is the primary source of risk for retail-oriented financial institutions. Information about past financial performance is the most critical signal that agricultural borrowers can send to distinguish their level of credit risk (B. Miller Ellinger & Lajili, 1993). However, data limitations are a major impediment in assessing farm financial performance (Zhang & Ellinger, 2006). Regarding small farmers, their business scale, geographic remoteness, informal accounting practices, and business and financial risks indicate high information needs to allow lenders to adequately manage credit risks (Barry & Robison, 2001).

Several studies have examined credit risk in agribusiness. A number of these studies used portfolio credit risk management models, seeking to estimate capital requirements for agricultural lenders. Katchova and Barry (2005) developed credit value-at-risk methods to calculate probability of default, loss given default, and expected and unexpected losses. Featherstone, Roessler, and Barry (2006) used credit scoring techniques to rate a portfolio of loans. Sherrick, Barry, and Ellinger (2000) and Dressler and Tauer (2016) developed credit risk valuation models for measuring credit risk to estimate expected and unexpected losses. Other studies have assessed the credit risks of individual loans through credit scoring models (Hou et al., 2005; L. H. Miller & LaDue, 1988; M. P. Novak et al., 1999; Turvey, 1991). However, the literature on credit scoring is very limited compared to the portfolio analysis literature (Thomas et al., 2017).

In regard to credit models that have been used for assessing the agricultural sector, those included are logistic regression (Durguner & Katchova, 2007; Hou et al., 2005; Limsombunchai et al., 2005; L. H. Miller & LaDue, 1988; M. P. Novak et al., 1999; Rambaldi et al., 1992; Römer, Römer, Musshoff, & Musshoff, 2017; Savitha, Savitha, Kumar K, & Kumar K, 2016), discriminant analysis (Bonazzi & Iotti, 2014; Rambaldi et al., 1992; Ziari et al., 1994) and machine learning techniques such as decision trees and neural networks (Limsombunchai et al., 2005; M. P. Novak et al., 1999).

Logistic regression is the classic and most widely used technique due to its simplicity and explanatory power (Siddiqi, 2017). Ziari et al. (1994) found that both mathematical programming techniques and statistical models performed equally well and that mixed integer-programming models perform better than parametric models. An advantage of non-parametric models is that

they can fit several distribution functions. Furthermore, when the data sample is small or if it is too dirty, non-parametric models such as neural networks may generate better results (Gustafson, Pederson, & Gloy, 2005).

Logistic regression is the technique most frequently applied in agricultural credit scoring (see table 1), with isolated studies showing a comparison of similar general linear classification techniques. Turvey (1991) used data from Canada's Farm Credit Corporation to compare the performance of four credit scoring models (linear probability model, discriminate analysis, logit, and probit) and found similar classification accuracies (between 71.5% and 67.1%) for these models. Non-linear techniques have also been benchmarked: Odeh, Featherstone, Sanjoy, et al. (2006) compared logistic regression, artificial neural networks and the adaptive neuro-fuzzy inference (ANFI) system to predict default using data from the Farm Credit System in the USA, identifying slight differences in prediction accuracies. ANFI gave better results than the other methods particularly in terms of sensitivity and specificity measures.

The types of variables used in the literature on credit scoring for farmers mainly describe financial ratios such as liquidity, profitability and leverage (Durguner & Katchova, 2007; Jouault & Featherstone, 2011; Ziari et al., 1994); farmer characteristics (educational level, age, goods etc.) (Limsombunchai et al., 2005); farm characteristics, such as types of crops and farm size (Limsombunchai et al., 2005; L. H. Miller & LaDue, 1988; M. P. Novak et al., 1999; Onyenucheya & Ukoha, 2007); and credit features, including credit history (Aruppillai & Phillip, 2014; Eyo & Ofem, 2014; Hou et al., 2005; Jouault & Featherstone, 2011). Other studies have used weather data (Pelka, Musshoff, & Weber, 2015; Römer et al., 2017) and variables related to the sustainability of crops (Henning & Jordaan, 2016). No studies have measured the relative impact of these sets of variables; each study apparently used what was available to them.

Turvey (1991) stressed the importance of including qualitative and quantitative attributes in credit scoring models. Gallagher (2001) indicated that a prediction model without non-financial variables could have model misspecification issues. Zech and Pederson (2003) identified the debt-to-asset ratio as a major predictor of repayment ability. Zech and Pederson (2003) also argued that both the total asset turnover ratio and family living expenses are strong predictors of the financial performance of a farm. Furthermore, it is a well-known fact that better sources of information are more useful in prediction than better models. This is discussed at length in Baesens, Roesch, and Scheule (2016) and shown empirically for newer so-called *alternative data* in a P2P and retail credit risk environment by Calabrese, Osmetti, and Zanin (2019) and Óskarsdóttir, Bravo, Sarraute, Vanthienen, and Baesens (2019).

In relation to the definition of default, the literature of credit scoring models for agribusiness takes different approaches. Jouault and Featherstone (2011) used the definition of 90 days past due, in concordance with the Basel Accords (Basel Committee on Banking Supervision, 2004). L. H. Miller and LaDue (1988) defined default as whenever a loan was refinanced. On the other hand, an alternative to traditional credit scoring is to use the coverage ratio directly as a measure

of creditworthiness (M. Novak & LaDue, 1994).

Regarding the purpose of the models, there are two categories: application scoring and behavioural scoring. The former is related to the decision whether to grant the loan, and the latter is about the decision on the credit limit or new product offers (when the credit is already granted). Most of the literature in credit scoring for agribusiness is related to the application scoring. L. H. Miller and LaDue (1988) evaluate existing borrowers using only financial ratios; their analysis did not use behavioural variables.

Table 1 presents a summary of previous work on lending in agribusiness, in terms of model types, variables used and the country in which the study was conducted. There are only a few analyses of all factors affecting the failure of farmers to repay, as most studies focus on the analysis of different types of models or variables, without considering the impact of the factors simultaneously. Limsombunchai et al. (2005), Eyo and Ofem (2014), and Savitha et al. (2016) analyse two different models and types of variables but do not take the size of the company and behavioural variables into account. This paper presents a simultaneous analysis of the impact of the creation of specialised variables (agribusiness and repayment behaviour), the type of classification techniques and company size. We perform this analysis to determine the most important factors when predicting the default of farmers debt and to make recommendations to agricultural lenders in relation to credit risk.

# 3 Financing Farmers in Developing Countries

According to Klein, Meyer, Hannig, Burnett, and Fiebig (2001), the types of rural lenders found in developing countries are the following:

- Formal lenders: banks, agricultural development agencies, rural branches of commercial banks, cooperative banks, rural banks/community banks.
- Semi-formal lenders: credit unions, other cooperatives, semi-formal local or community banks, NGOs.
- Informal lenders: relatives and friends, independent moneylenders, rotating savings and credit associations.
- Credit interconnected systems: suppliers of agricultural inputs/crop buyers, agro-industries.

The sources of formal financing, such as commercial banks, have a strong aversion to lending to small farmers because of the characteristics of this sector, with relatively higher and complex risk profiles (ODEPA, 2009). Other sources of funding, particularly interconnected systems (suppliers of agricultural inputs/crop buyers), "have an advantage in relation to customer closeness and

Table 1: Credit Scoring Models for Farmers. The models that were applied were: logistic regression (LR), multinominal logistic regression (MLR), discriminant analysis (DA), variations of discriminant analysis (MDLA, LDA and FLDA), recursive partitioning algorithm (RPA) equivalent to decision trees and regression models (RM)

| **Author (Year)** | **Models** | **Variables** | **Country** |
|---|---|---|---|
| Miller and LaDue (1988) | LR | Farm size, liquidity, solvency, profitability, capital efficiency, operating efficiency. | USA |
| Rambaldi *et al.* (1992) | DA, LR | Liquidity, debt utilisation, profitability, assets, operational efficiency. | USA |
| Ziari *et al.* (1994) | DA, FLDA, LDA | Financial ratios. | Canada |
| Novak *et al.* (1999) | RPA, LR | Debt-to-asset ratio, current ratio. | USA |
| Hou *et al.*(2005) | LR | Demographic statistics, business and loan information. | USA |
| Limsombunchai *et al.* (2005) | LR, ANN | Borrower characteristics, credit risk proxies, relationship indicators. | Thailand |
| Durguner and Katchova (2007) | LR | Financial ratios. | USA |
| Onyenucheya and Ukoha (2007) | RM, DA | Farmer characteristics, credit features, ratios, distance (home - loan source). | Nigeria |
| Jouault and Featherstone (2011) | LR | Ratios, credit information. | France |
| Eyo and Ofem (2014) | DA, RM | Borrower features, loan information, financial ratios, farm size. | Nigeria |
| Aruppillai and Phillip (2014) | RM | Borrower features, loan information. | Sri Lanka |
| Bonazzi and Iotti (2014) | MDLA | Financial ratios. | Italy |
| Savitha *et al.* (2016) | LR, MLR | Borrower characteristics (both financial and non-financial) and relationship indicators. | India |
| Römer (2017) | LR | Socio-economic characteristics of clients, financial ratios, credit related, working experience. | Madagascar |

Table 2: Number of Farmer Loans in Chile. Original data from ODEPA (2013)

| **Source** | **Amount (mln USD)** | **Share** |
|---|---|---|
| INDAP (Agricultural Development Institute) | 69.81 | 1.1% |
| Input suppliers | 711.16 | 11.6% |
| Agriculture contract | 68.90 | 1.1% |
| Commodity exchange | 53.87 | 0.9% |
| Foreign investment | 39.51 | 0.6% |
| Credit unions | 11.35 | 0.2% |
| Factoring | 4.17 | 0.1% |
| **Subtotal** | **958.77** | **15.6%** |
| Banks | 5,192.60 | 84.4% |
| **Total** | **6,151.37** | **100.0%** |

knowledge of different fields, attributes that are valued beyond the rate interest charge" (ODEPA, 2013).

In the source country of our data, 17.9% of farmers use some form of credit to finance their business (EME, 2014). Table 2 shows the sources of financing used by these farmers (ODEPA, 2013). Most of the farmers chose bank credits (84.4%), with the second most important source of financing corresponding to suppliers of agricultural inputs (11.6%).

Using data from farmers seeking loans in credit interconnected systems can permit the determination of relevant factors in this segment, with reference to their repayment behaviour. This is due to the knowledge of the agricultural area and the proximity of these institutions to their customers.

# 4 Data

This chapter describes the dataset used in this analysis. In particular, we provide details on the data preparation, and we present the variables used in the scorecard. Moreover, we explain the transformations applied to the data.

## 4.1 Data Preparation

We used data provided by a Chilean company that grants loans to farmers for the supply of inputs, besides providing support services. The data were anonymised to protect customer confidentiality and identity. The data relate to 6,658 customers who were approved between January 2007 and December 2013. The data include a subset of the customers' application characteristics and full subsequent repayment behaviour up to December 2014. We considered a sample of 161,613 credit sales, splitting the dataset into three segments. The person (independent farmers) segment has 48,875 cases; the company segment has 58,443 cases; and the holding segment has 54,295 cases. The default percentage across the sample is 2.55%, and the rates by segment are 2.56%, 2.48%,

and 2.64%, for persons, companies and holdings respectively.

The data time period reflects an entire economic cycle, including the end of an economic expansion, a recession and a recovery; thus, given that our objective is to study the impacts of our factors on modelling agricultural loans, we consider the data both sufficient to cover the application of these technologies under most conditions and robust to changes in the economic conditions. We can possibly extrapolate this to multiple countries, as Chile is an upper-middle-income country with very large holding corporations (represented in the holdings dataset) that are more competitive than many companies from high-income countries (The World Economic Forum, 2017). We also study small farmers with a reality much closer to a low-to-middle-income country, particularly those within the supply chains of the large companies (Reardon, Barrett, Berdegué, & Swinnen, 2009). The studies of small and medium agribusinesses lie somewhere in the middle; they are much more representative of the Chilean upper-middle-income reality, given that they are much more dependent on the local economy than large holding companies. Thus, we believe our dataset and segmentation both create an interesting profile of the use of models for risk management in the agribusiness and represent different conditions and realities worldwide.

The best practice, according to the literature (Siddiqi, 2007), is to consider default as occurring when one payment is more than 90 days in arrears during the first 12 months after granting the loan. We use the same definition for this study. The 90-days definition of the target variable corresponds to the definition of a good/bad complaint within the Basel Accords (Basel Committee on Banking Supervision, 2004), which considers an obligor bad if the bank determines that the obligor is unlikely to pay its credit obligations or if any material credit obligation is past due by more than 90 days. The definition of default can be applied at the level of a loan (a particular facility) for retail exposures, that is, a default by a borrower on one loan does not imply that all other loans are in default (Basel Committee on Banking Supervision, 2004). In this sense, the definition is applied at loan level because most of the company's loans can be classified as retail exposures, especially the loans of persons and small companies.

Given that some borrowers have a history with the company, we also need to study past behaviour during a set period of time. This requires setting up a "Performance Window" during which each loan is studied, a period that again is usually considered to be from 6 to 12 months. Considering the periodicity of crops, a 12-month performance window gives the best chance of capturing the borrowers' behaviour, by capturing an entire period.

We do not add macroeconomic variables in this study because the initial idea of the model was to first consider the standard approach of estimating scores with no macroeconomic variables in an unconditional model, and then to calibrate this model over macro variables for provisioning and capital requirements in the IFRS 9's expected credit loss framework. In this framework, the probability of default can be obtained by using internal historical data adjusted by forward-looking information according to different possible macroeconomic scenarios.

In terms of data preprocessing, we removed variables with low variability (if more than 95% of

the observations showed the same category) and with more than 30% of missing values.

The variables selected for this study fall into the following categories:

- Socio-demographic variables: the region of the borrower's residence; the economic sector in which the farmer operates, according to company's internal classification (agricultural and others); the level of purchases made during the last year and the type of client (person, company or holding company).
- Agribusiness variables: the reported income of the borrower, the cost of operation, types of crops (cherry, plum, corn, apple, walnut, meadows, wheat, wine grapes and others) and information about the customer's properties (related to location, plantation area and number of properties).
- Credit variables: the attributes of the loan and the history of the customer in the company (for example, the client's length of tenure, the branch office region of credit application, instalment and loan amount, payment type or term type according to payment frequency of the loan).
- Behavioural variables related to payment behaviour, which can be divided into three time windows: the last 3 months, the period of the last 3 to 6 months and the period of the last 6 to 9 months. As the values of behavioural variables change over the performance window, we computed the maximum, the minimum, the average, and the number of increments and decrements in the standing balance and various ratios, such as amounts of arrears and days in arrears.

In total, the dataset is composed of 5 socio-demographic variables, 17 agribusiness-related variables, 19 credit variables and 42 behavioural variables.

## 4.2 Variable Selection and Transformation

The variable selection process was developed in two stages. To test the independence of the explanatory variables with the target variable, we used the $\chi^2$ test for categorical variables and the Kolmogorov-Smirnov test for continuous variables. We removed the variables that did not show a relationship with the target variable at a 95% confidence level. Afterwards, we created clusters of the independent variables in order to reduce the dimensionality of the dataset using the *ClustOfVar* algorithm (Brida, Fasone, Scuderi, & Zapata-Aguirre, 2014). This algorithm applies K-means clustering to categorical and continuous variables using a synthetic variable calculated by principal component analysis as a centre (Kiers, 1991).

We also performed a multicollinearity analysis by removing variables with a variance inflation factor higher than 5 (Mansfield & Helms, 1982). Finally, we used a stepwise selection procedure,

and we removed the variables that had a significance level higher than 0.05 in each iteration. We finally obtained 30 variables for the whole sample, 33 variables for the data on individuals, 32 variables for the companies and 29 variables for holding companies.

To normalise the dataset and centre it using a common scale, we applied the weight of evidence (WOE) transformation to the variables, computed as follows:

$$WOE_{c_v,v} = ln\left(\frac{DistrGood_{c_v,v}}{DistrBad_{c_v,v}}\right), \tag{1}$$

where $v$ is the index of the variables that are available, and $c_v$ is the index of each variable's categories. $DistrGood_{c_v,v}$ and $DistrBad_{c_v,v}$ are the proportions of cases of the attribute that belong to the good and bad classes respectively, over the total cases of the class. We used this transformation because it is a common procedure in credit scoring models (Siddiqi, 2007). To apply this transformation to the continuous variables, we discretised them using classification trees. For the categorical variables, we aggregated categories in order to have at least 5% of the total cases in each category.

The resulting dataset is clean of outliers, centred and discretised to better capture behaviour. We now proceed with the experimental design to test our hypotheses.

# 5 Experimental Design

The experimental design of this study consists of a factorial experimental setup to assess the effects of three different factors on the performance of default prediction for farmers. The first factor represents the type of explanatory variables and consists of four possible levels given by the credit variables, the behavioural variables, the socio-demographic variables and the agribusiness variables. This factor both reflects the amount of information that a company must store and supports the complexity analysis, since more complex patterns require more data; it also allows us to study the diversity of these patterns. If more data sources are needed, it suggests that a mix of different risks affects the ability of borrowers to satisfy their obligations.

Formally, let $\mathbf{x}$ be the set of all the independent variables; let $x_{ag}$ be the subset of agribusiness variables, $x_{sd}$ the subset of sociodemographic variables, $x_{ap}$ the subset of credit variables and $x_{bh}$ the subset of behavioural variables. We estimate the probability of default $P(y = 1|\mathbf{x})$ as a function of the four subsets of variables:

$$P(y = 1|\mathbf{x}) = f(x_{ag}, x_{ap}, x_{sd}, x_{bh}). \tag{2}$$

The second factor of the experimental design concerns the classification techniques. It has three possible levels given by logistic regression, random forest and neural network analysis. The main question to be answered by this factor is the relevance of complex, non-linear patterns in

the behaviour of the borrowers. If a more complex model results in a much higher discrimination capacity, then we can conclude that there is a much more complex structure among the borrowers' behaviour, which impacts the ability for small lenders to model risk effectively.

The first model selected is given by logistic regression, a widely used approach in credit risk analysis (Baesens, Van Gestel, et al., 2003). This model is the basic generalised linear model and can correctly represent the relationship between linear combinations of variables in the sample and the logit or the logarithm of the odds that a borrower presents the event being studied (Hosmer & Lemeshow, 2000). Formally, the logistic regression models the probability of the event as

$$f(x) = \frac{1}{1 + exp(-\beta^t \cdot x)}, \tag{3}$$

where $x = (x_{ag}, x_{ap}, x_{sd}, x_{bh})$, the vector of the variables in the model, and $\beta$ is the vector of weights each variable has in the model.

We also use the neural network, a powerful non-linear but difficult-to-interpret model (Hassoun, 1995) that can capture a more complex non-linear structure in a single expression. We use a shallow model representation, which is effective at looking at general non-linear patterns in the data. We use one hidden layer and sigmoid transfer and output functions in the architecture. The number of neurons in the hidden layer is obtained by maximising the area under the ROC curve (AUC) in a validation set.

The last approach is the random forest method, which is a robust alternative for predicting default due to its ability to detect complex patterns following a deep analysis of all the subsets of the input space (Breiman, 2001). The random forest approach combines decision trees so that they all use a separate sample of cases and variables simultaneously, producing diverse trees that create, when evaluated jointly, a very detailed analysis of the input space (deep search). For each tree, a bootstrapped sample of the data, usually of size 64.2%, and a sample of the variables, usually 1/3, is selected to train it. Assuming that each tree produces a binary output given by $o_i$, we can generate a valid output by simply averaging each individual tree. As shown in Probst and Boulesteix (2018), the best strategy for selecting the number of trees is to simply train as many as possible. We chose a number of trees such that no improvement occurred in the out-of-bag sample when adding a new one.

In previous studies, these three methods have been identified as the most accurate for building credit scorecards for each level of complexity (Lessmann, Seow, Baesens, & Thomas, 2013). The chosen models are in very different areas of the interpretability/complexity spectrum. A logistic regression will only account for linear relationships between variables but will provide a very clear picture of the way the variables have an effect on the target, in terms of both magnitude of the impact and its direction, that is, if a larger value of a given variable implies an increase or decrease in the borrower's risk. A random forest is exactly at the other extreme, because the only information that can be extracted is the contribution of each variable, which is done by comparing

metrics (usually AUC or accuracy) between trees that include a certain variable and those that do not. Neural networks lie somewhere in between because they are a model represented by a unique function – as opposed to random forests, which are ensembles of decision trees – and it is possible to extract rules from their output (Baesens, Setiono, Mues, & Vanthienen, 2003); otherwise, they are black box models. These three models give a very broad picture of the technological abilities that are currently available for extracting patterns for structured data, and they allow us to profile the usefulness of complex models versus simpler solutions.

The third factor represents the type of clients over three possible levels given by company, holding company or person. Person refers to customers who apply for credit individually and are not associated with or do not belong to any company. The remaining categories – enterprises and holding companies – are clients who represent a company or a business organisation that controls a number of companies. This factor illuminates not only the differences that arise from multiple organisational structures but also how their composition, from single farmers operating on their own to large holdings, affects the lender's ability to capture credit risk through a statistical procedure. If each segment is completely different, then more scorecards and therefore more independent systems need to be kept in parallel to serve the customers. This again has managerial implications because the lender has a more complex risk area, thereby increasing the cost of sustaining proper operations.

For each possible combination, we estimated a scoring model and computed the AUC, a common index reported in the literature for analysing predictive accuracy (Lobo, Jiménez-Valverde, & Real, 2008) and for comparing the predictive capabilities of the model. In the next section, we consider all the possible combinations given by a total number (135) of 15x3x3 models.

# 6 Empirical Results

This section presents the main results of the study in relation to the analysis of the impact of various factors – explanatory variables, modelling techniques and segmentation – in default prediction in the agribusiness sector.

## 6.1 Explanatory Variables

After applying the WOE transformation, we analysed the ability of the explanatory variables to predict the good and bad cases. We determined whether the relationship of the independent variable coheres with expectations. For each variable, the information value (IV) was computed, a measure that comes from information theory (Kullback, 1997) and that reveals the predictive power of the attributes. According to Siddiqi (2007), a variable is highly predictive if its information value is greater than 0.3. The results for all the customers are presented in table 3. We use the denotations "sd", "ap", "ag" and "bh" for the socio-demographic variables, the credit

variables, the agribusiness variables and the behavioural variables respectively. We also report the strength of the relationship between each explanatory variable and the dependent variable in table 3 following Siddiqi (2007). The variables ArrearsLast3M, Arrears3to6Months, TimelyPayLast3M, CropTypeG2 and TimelyInstLast3M, belonging to the behavioural and agribusiness groups show the higher information values. These results show that the behavioural variables represent the strongest predictors of capacity to repay, as is the case in consumer lending. The signal given by the most recent payment behaviour (previous 3-6 months) is of greater relevance within this subset. A more important variable for this segment, highly ranked in the sample and with very strong explanatory power, is the type of crop. This indicates that the seasonality of crops will be a very strong indicator of future performance, but at the same time the inclusion of this variable brings the risk that the model can be affected by an external impact on the crops (for example, a particular climate event); thus, the predictive capability of the model might be affected. Usually, a recalibration using more recent data is all that is needed to recover from this circumstance, so this risk should not discourage a potential user from including the variable.

## 6.2 Predictive Accuracy

Because the datasets were imbalanced with respect to the classes of the target variable, we applied the synthetic minority over-sampling technique (SMOTE), a method that combines over-sampling and undersampling to generate balanced datasets (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). To avoid overfitting, we estimated the models and validated them both on an out-of-sample set, generated by randomly drawing 30% of the customers, and on an out-of-time sample, given by the credit sales after January 2014.

Both neural networks and random forest require tuning certain parameters in order to find the choices that better represent the patterns in the data. Neural networks require the number of neurons in the hidden layer and the number of training epochs, while random forest requires the maximum depth per tree and the number of variables per tree. These parameters are adjusted by grid search, finding the optimal parameter for each of the 135 models using 20% of the training sample.

To measure the predictive accuracy, we used AUC, which is the area under a receiver operating characteristic (ROC) curve. This curve corresponds to the plotted values of the probability of true positives (correctly predicted defaults) and the probability of false positives values (incorrectly predicted good loans), illustrating a trade-off between the captured response fraction and the false positive fraction. Each point on the ROC chart corresponds to a specific fraction of cases, ranked by their predicted value. AUC is the probability that a randomly chosen positive case is correctly rated, having a greater suspicion than a randomly chosen negative case (Hanley & McNeil, 1982).

The AUCs calculated for all combinations are reported in tables 4, 5, 6 and 7. If we constrain each model to include only one type of variable, then behavioural variables, followed by

agribusiness-related characteristics, show the best performance. In most of the combinations, the highest accuracy is achieved using information from different groups of variables. For example, table 4 shows that the higher AUC on all customers for each modelling technique is achieved by using all the explanatory variables on the out-of-sample test set.

To measure the contribution of each group of variables in terms of performance, we computed the normalised AUC, dividing the AUC by the maximum out-of-time AUC for each segment. We show the results in tables 8, 9, 10 and 11. In general, behavioural variables increase the AUCs from 5% to 20%, whereas agribusiness variables contribute from 5% to 10% in extra predictive capability. In particular, behavioural variables show the highest impact on the AUC for all the customers in a logistic regression model. We obtained similar results for all the other segments of customers.

Applying segmentation of the customers can increase the AUC by up to 2.7% on the out-of-sample data. Conversely, the accuracy decreases by 2.5% if the segmentation is implemented on the out-of-time sample, indicating that using a one-size-fits-all model can deliver a more stable result. Tables 8, 9, 10 and 11 show that the best model with all the available variables' types is the random forest, followed by the logistic regression approach. The neural network model shows the worst performance on the out-of-time sample for all the customers, as is also displayed in figure 1.

On the another hand, in order to check the prediction stability for each of the applied techniques, using the models that consider all the types of variables, we plotted the predicted default rate of the models versus the real one. To check how the crop periodicity influences the outcome, we used an out-of-time sample of 1 year. Based on a profitability criterion, namely the expected maximum profit measure (Verbraken, Bravo, Weber, & Baesens, 2014), we obtained the optimal cut-off point for each model.

The results can be seen in figure 2. In general, the three applied techniques were able to capture the default rate periodicity. Random forest was the technique with the best performance during all periods, with a default rate that is closer to the real default rate.

To sum up, logistic regression performs well in predictive accuracy compared to machine learning techniques (random forests and neural networks). On the other hand, neural networks have good performance in out-of-sample and unstable results in samples out-of-time. Random forests are significantly better in the sample out-of-time; this can be explained by the fact that random forests use multiple decision trees and different samples and variables to generate robust results and avoid overfitting.

Table 3: Information Values (IV) and the Strength of the Relationship for Each Explanatory Variable on All Customer Segments

| Variable | Description | Group [a] | IV | Strength |
|---|---|---|---|---|
| ArrearsLast3M | Avg. days in arrears in last 3 months | bh | 0.56 | Strong |
| Arrears3to6Months | Avg. days in arrears in last 3–6 months | bh | 0.44 | Strong |
| TimelyPayLast3M | Avg. amount paid on time and total paid in last 3 months | bh | 0.30 | Strong |
| CropTypeG2 | Crop type | ag | 0.29 | Strong |
| TimelyInstLast3M | Avg. ratio between payment and instalment in last 3 months | bh | 0.29 | Strong |
| TotalBalance | Total amount owed | ap | 0.24 | Strong |
| TimelyPay3Mto6M | Avg. of ratio of amount paid on time and total paid in last 3–6 months | bh | 0.21 | Strong |
| TimelyPay6Mto9M | Avg. of ratio of amount paid on time and total paid in last 6–9 months | bh | 0.21 | Strong |
| TimelyInstLast3M | Avg. ratio between payment and instalment in last 3 months | bh | 0.20 | Strong |
| NrTimelyLast3M | No. of instalments paid on time in last 6–9 months | bh | 0.20 | Medium |
| RegionG1 | Geographic region | sd | 0.18 | Medium |
| Cost | Agricultural investment | ag | 0.17 | Medium |
| LevelPurchases | Purchases level | sd | 0.13 | Medium |
| IncomeHectare | Income per hectare | ag | 0.13 | Medium |
| CostProperty | Ratio between agricultural investment and no. of properties of the customer | ag | 0.13 | Medium |
| AmountArrears3Mto6M | Avg. arrears amount in last 6–9 months | bh | 0.12 | Medium |
| CropsNumber | No. of different crop types | ag | 0.12 | Medium |
| Income | Agricultural activity income | ag | 0.11 | Medium |
| AmountArrears6Mto9M | Avg. arrears amount in last 6–9 months | bh | 0.11 | Medium |
| ArrearsIncreaseLast3M | No. of increases of the arrears amount in last 3 months | bh | 0.10 | Medium |
| CostHectare | Agricultural activity cost per hectare | ag | 0.10 | Medium |
| NrPastDueLast3M | No. of past due instalments in last 3–6 months | bh | 0.10 | Medium |
| ArrearsIncrease3Mto6M | No. of increases of the arrears amount in last 3-6 months | bh | 0.09 | Weak |
| Tenure | If the credit applicant is a client | ap | 0.09 | Weak |
| PropertyLocationN | No. of different property locations | ag | 0.08 | Weak |
| PropertyDistance | Avg. distance between each property and its nearest branch office | ag | 0.07 | Weak |
| PreviousPurchasesN | No. of previous purchases | ap | 0.06 | Weak |
| NrPastDue3Mto6M | No. of past due instalments in last 6–9 months | bh | 0.04 | Weak |
| ArrearsDecrease3Mto6M | No. of decreases in arrears amount in last 3–6 months | bh | 0.02 | Weak |
| ArrearsDecrease6Mto9M | No. of decreases in arrears amount in last 6–9 months | bh | 0.02 | Weak |
| TimeLastMaturity | Months since the most recent maturity | ap | 0.01 | Unpredictive |

[a] The abbreviations *sd*, *ap*, *ag* and *bh* stand for sociodemographic variables, the credit variables, the agribusiness variables and the behavioural variables, respectively.

Table 4: AUC Indices for all Borrowers

| **Variables** [a] | **Logistic Regression** | | **Neural Networks** | | **Random Forests** | |
|---|---|---|---|---|---|---|
| | Out of sample | Out of time | Out of sample | Out of time | Out of sample | Out of time |
| sd | 0.650 | 0.595 | 0.648 | 0.609 | 0.630 | 0.589 |
| ag | 0.675 | 0.663 | 0.767 | 0.668 | 0.717 | 0.633 |
| ap | 0.692 | 0.686 | 0.707 | 0.689 | 0.695 | 0.689 |
| bh | 0.736 | 0.806 | 0.819 | 0.761 | 0.762 | 0.797 |
| sd + ag | 0.706 | 0.675 | 0.820 | 0.693 | 0.818 | 0.720 |
| sd + ap | 0.714 | 0.681 | 0.746 | 0.711 | 0.726 | 0.702 |
| sd + bh | 0.756 | 0.802 | 0.828 | 0.783 | 0.842 | 0.830 |
| ag + ap | 0.733 | 0.720 | 0.823 | 0.727 | 0.820 | 0.743 |
| ag + bh | 0.779 | 0.816 | 0.854 | 0.769 | 0.882 | 0.846 |
| ap + bh | 0.779 | 0.821 | 0.848 | 0.773 | 0.840 | 0.830 |
| sd + ag + ap | 0.745 | 0.716 | 0.847 | 0.720 | 0.870 | 0.774 |
| sd + ag + bh | 0.786 | 0.813 | 0.869 | 0.774 | 0.902 | 0.865 |
| sd + ap + bh | 0.785 | 0.816 | 0.851 | **0.784** | 0.873 | 0.844 |
| ag + ap + bh | 0.800 | **0.828** | 0.861 | 0.762 | 0.899 | 0.871 |
| sd + ag + ap + bh | **0.803** | 0.824 | **0.871** | 0.783 | **0.917** | **0.879** |

[a] The abbreviations *sd*, *ap*, *ag* and *bh* stand for sociodemographic variables, the credit variables, the agribusiness variables and the behavioural variables, respectively.

Table 5: AUC Indices for the Subset of Persons

| Variables [a] | Logistic Regression | | Neural Networks | | Random Forests | |
|---|---|---|---|---|---|---|
| | Out of sample | Out of time | Out of sample | Out of time | Out of sample | Out of time |
| sd | 0.556 | 0.640 | 0.662 | 0.559 | 0.643 | 0.635 |
| ag | 0.740 | 0.731 | 0.806 | 0.673 | 0.812 | 0.694 |
| ap | 0.792 | 0.738 | 0.783 | 0.683 | 0.800 | 0.671 |
| bh | 0.736 | 0.774 | 0.821 | 0.776 | 0.759 | 0.796 |
| sd + ag | 0.741 | 0.737 | 0.861 | 0.735 | 0.886 | 0.803 |
| sd + ap | 0.798 | 0.760 | 0.820 | 0.694 | 0.834 | 0.748 |
| sd + bh | 0.734 | 0.784 | 0.855 | 0.771 | 0.836 | 0.735 |
| ag + ap | 0.819 | 0.770 | 0.832 | 0.666 | 0.909 | 0.796 |
| ag + bh | 0.807 | 0.820 | 0.863 | 0.779 | 0.896 | 0.785 |
| ap + bh | 0.820 | 0.795 | 0.880 | **0.837** | 0.898 | 0.869 |
| sd + ag + ap | 0.820 | 0.773 | 0.836 | 0.705 | 0.924 | 0.788 |
| sd + ag + bh | 0.806 | 0.819 | 0.905 | 0.770 | 0.919 | 0.774 |
| sd + ap + bh | 0.820 | 0.799 | **0.905** | 0.799 | 0.907 | **0.854** |
| ag + ap + bh | 0.845 | 0.822 | 0.897 | 0.775 | 0.933 | 0.844 |
| sd + ag + ap + bh | **0.845** | **0.826** | 0.898 | 0.813 | **0.938** | 0.833 |

[a] The abbreviations *sd*, *ap*, *ag* and *bh* stand for sociodemographic variables, the credit variables, the agribusiness variables and the behavioural variables, respectively.

Table 6: AUC Indices for the Subset of Companies

| Variables [a] | Logistic Regression | | Neural Networks | | Random Forests | |
|---|---|---|---|---|---|---|
| | Out of sample | Out of time | Out of sample | Out of time | Out of sample | Out of time |
| sd | 0.681 | 0.571 | 0.697 | 0.569 | 0.680 | 0.572 |
| ag | 0.699 | 0.646 | 0.819 | 0.746 | 0.791 | 0.734 |
| ap | 0.737 | 0.636 | 0.785 | 0.638 | 0.762 | 0.622 |
| bh | 0.733 | 0.798 | 0.816 | 0.745 | 0.788 | 0.793 |
| sd + ag | 0.726 | 0.653 | 0.881 | 0.805 | 0.884 | 0.796 |
| sd + ap | 0.755 | 0.637 | 0.834 | 0.646 | 0.827 | 0.675 |
| sd + bh | 0.771 | 0.782 | 0.876 | 0.758 | 0.875 | 0.812 |
| ag + ap | 0.769 | 0.713 | 0.891 | 0.708 | 0.891 | 0.769 |
| ag + bh | 0.789 | 0.805 | 0.902 | 0.776 | 0.907 | 0.825 |
| ap + bh | 0.810 | 0.799 | 0.872 | 0.732 | 0.883 | 0.821 |
| sd + ag + ap | 0.777 | 0.708 | 0.886 | 0.682 | 0.917 | 0.840 |
| sd + ag + bh | 0.798 | 0.807 | 0.896 | 0.772 | 0.933 | 0.848 |
| sd + ap + bh | 0.813 | 0.795 | 0.867 | 0.722 | 0.907 | 0.830 |
| ag + ap + bh | 0.825 | **0.822** | 0.913 | 0.756 | 0.931 | 0.853 |
| sd + ag + ap + bh | **0.826** | 0.821 | **0.925** | **0.830** | **0.939** | **0.870** |

[a] The abbreviations *sd*, *ap*, *ag* and *bh* stand for sociodemographic variables, the credit variables, the agribusiness variables and the behavioural variables, respectively.

Table 7: AUC Indices for the Subset of Holding Companies.

| Variables [a] | Logistic Regression | | Neural Network | | Random Forest | |
|---|---|---|---|---|---|---|
| | Out of Sample | Out of time | Test | Out of time | Out of sample | Out of time |
| sd | 0.654 | 0.574 | 0.656 | 0.575 | 0.654 | 0.574 |
| ag | 0.749 | 0.616 | 0.835 | 0.711 | 0.809 | 0.687 |
| ap | 0.705 | 0.599 | 0.724 | 0.581 | 0.760 | 0.598 |
| bh | 0.778 | 0.807 | 0.847 | 0.771 | 0.801 | 0.813 |
| sd + ag | 0.770 | 0.628 | 0.900 | 0.767 | 0.874 | 0.748 |
| sd + ap | 0.738 | 0.625 | 0.834 | 0.712 | 0.853 | 0.708 |
| sd + bh | 0.789 | 0.802 | 0.883 | 0.762 | 0.847 | 0.802 |
| ag + ap | 0.778 | 0.642 | 0.907 | 0.722 | 0.920 | 0.789 |
| ag + bh | 0.826 | 0.784 | **0.911** | **0.788** | 0.907 | 0.818 |
| ap + bh | 0.810 | **0.816** | 0.883 | 0.714 | 0.889 | 0.816 |
| sd + ag + ap | 0.792 | 0.655 | 0.880 | 0.677 | 0.928 | 0.799 |
| sd + ag + bh | 0.832 | 0.782 | 0.875 | 0.717 | 0.929 | 0.825 |
| sd + ap + bh | 0.817 | 0.810 | 0.905 | 0.736 | 0.922 | 0.828 |
| ag + ap + bh | 0.842 | 0.802 | 0.852 | 0.689 | 0.937 | 0.852 |
| sd + ag + ap + bh | **0.847** | 0.798 | 0.858 | 0.634 | **0.944** | **0.863** |

[a] The abbreviations *sd*, *ap*, *ag* and *bh* stand for sociodemographic variables, the credit variables, the agribusiness variables and the behavioural variables, respectively.

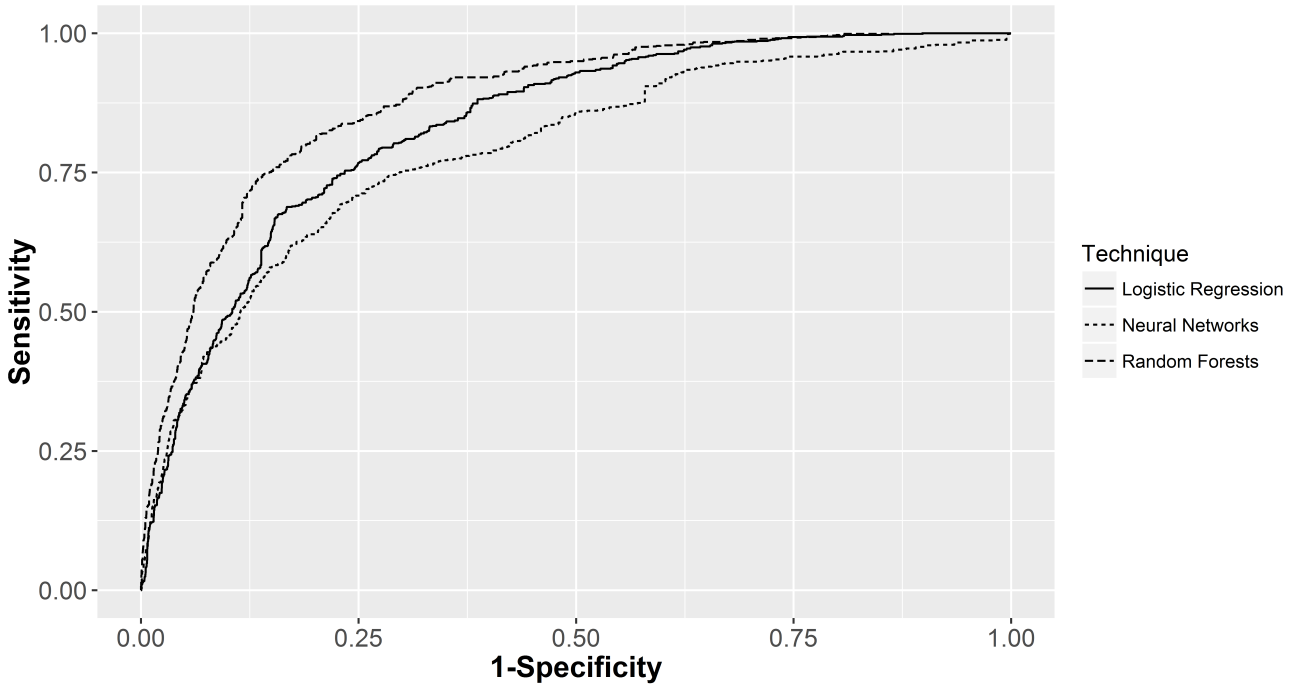


Figure 1: The ROC curves of all the customers on an out-of-time sample for different classification techniques.

Table 8: Normalised AUC Indices, All Customers, Out-of-Time Sample

| **Variables** [a] | **Logistic Regression** | **Neural Network** | **Random Forest** |
|---|---|---|---|
| sd | 0.677 | 0.693 | 0.670 |
| ag | 0.754 | 0.760 | 0.720 |
| ap | 0.781 | 0.784 | 0.783 |
| bh | 0.917 | 0.866 | 0.906 |
| sd + ag | 0.767 | 0.788 | 0.819 |
| sd + ap | 0.775 | 0.809 | 0.799 |
| sd + bh | 0.912 | 0.890 | 0.944 |
| ag + ap | 0.819 | 0.827 | 0.846 |
| ag + bh | 0.928 | 0.875 | 0.963 |
| ap + bh | 0.935 | 0.879 | 0.945 |
| sd + ag + ap | 0.815 | 0.819 | 0.881 |
| sd + ag + bh | 0.925 | 0.881 | 0.984 |
| sd + ap + bh | 0.928 | 0.892 | 0.961 |
| ag + ap + bh | 0.942 | 0.867 | 0.991 |
| sd + ag + ap + bh | 0.937 | 0.891 | 1.000 |

[a] The abbreviations *sd*, *ap*, *ag* and *bh* stand for sociodemographic variables, the credit variables, the agribusiness variables and the behavioural variables, respectively.

Table 9: Normalised AUC Indices, Persons, Out-of-Time Sample

| **Variables** [a] | **Logistic Regression** | **Neural Network** | **Random Forest** |
|---|---|---|---|
| sd | 0.736 | 0.643 | 0.730 |
| ag | 0.841 | 0.774 | 0.798 |
| ap | 0.849 | 0.786 | 0.772 |
| bh | 0.890 | 0.893 | 0.916 |
| sd + ag | 0.848 | 0.846 | 0.923 |
| sd + ap | 0.875 | 0.799 | 0.860 |
| sd + bh | 0.902 | 0.887 | 0.845 |
| ag + ap | 0.885 | 0.766 | 0.916 |
| ag + bh | 0.943 | 0.896 | 0.903 |
| ap + bh | 0.914 | 0.963 | 1.000 |
| sd + ag + ap | 0.889 | 0.811 | 0.907 |
| sd + ag + bh | 0.942 | 0.886 | 0.891 |
| sd + ap + bh | 0.919 | 0.919 | 0.982 |
| ag + ap + bh | 0.945 | 0.892 | 0.971 |
| sd + ag + ap + bh | 0.951 | 0.935 | 0.958 |

[a] The abbreviations *sd*, *ap*, *ag* and *bh* stand for sociodemographic variables, the credit variables, the agribusiness variables and the behavioural variables, respectively.

Table 10: Normalised AUC Indices, Companies, Out-of-Time Sample

| **Variables** [a] | **Logistic Regression** | **Neural Network** | **Random Forest** |
|---|---|---|---|
| sd | 0.656 | 0.653 | 0.658 |
| ag | 0.742 | 0.857 | 0.843 |
| ap | 0.731 | 0.733 | 0.715 |
| bh | 0.917 | 0.857 | 0.911 |
| sd + ag | 0.750 | 0.925 | 0.915 |
| sd + ap | 0.732 | 0.742 | 0.776 |
| sd + bh | 0.898 | 0.871 | 0.934 |
| ag + ap | 0.819 | 0.813 | 0.884 |
| ag + bh | 0.925 | 0.892 | 0.948 |
| ap + bh | 0.919 | 0.841 | 0.944 |
| sd + ag + ap | 0.813 | 0.784 | 0.965 |
| sd + ag + bh | 0.928 | 0.887 | 0.974 |
| sd + ap + bh | 0.913 | 0.829 | 0.954 |
| ag + ap + bh | 0.944 | 0.869 | 0.980 |
| sd + ag + ap + bh | 0.943 | 0.954 | 1.000 |

[a] The abbreviations *sd*, *ap*, *ag* and *bh* stand for sociodemographic variables, the credit variables, the agribusiness variables and the behavioural variables, respectively.

Table 11: Normalised AUC Indices, Holding companies, Out-of-time sample

| **Variables**[a] | **Logistic Regression** | **Neural Network** | **Random Forest** |
|---|---|---|---|
| sd | 0.665 | 0.666 | 0.665 |
| ag | 0.714 | 0.824 | 0.796 |
| ap | 0.694 | 0.674 | 0.693 |
| bh | 0.935 | 0.893 | 0.942 |
| sd + ag | 0.728 | 0.889 | 0.867 |
| sd + ap | 0.724 | 0.824 | 0.821 |
| sd + bh | 0.930 | 0.883 | 0.929 |
| ag + ap | 0.744 | 0.836 | 0.914 |
| ag + bh | 0.908 | 0.913 | 0.948 |
| ap + bh | 0.946 | 0.828 | 0.945 |
| sd + ag + ap | 0.759 | 0.784 | 0.926 |
| sd + ag + bh | 0.906 | 0.830 | 0.956 |
| sd + ap + bh | 0.939 | 0.853 | 0.960 |
| ag + ap + bh | 0.929 | 0.798 | 0.987 |
| sd + ag + ap + bh | 0.924 | 0.734 | 1.000 |

[a] The abbreviations *sd*, *ap*, *ag* and *bh* stand for sociodemographic variables, the credit variables, the agribusiness variables and the behavioural variables, respectively.

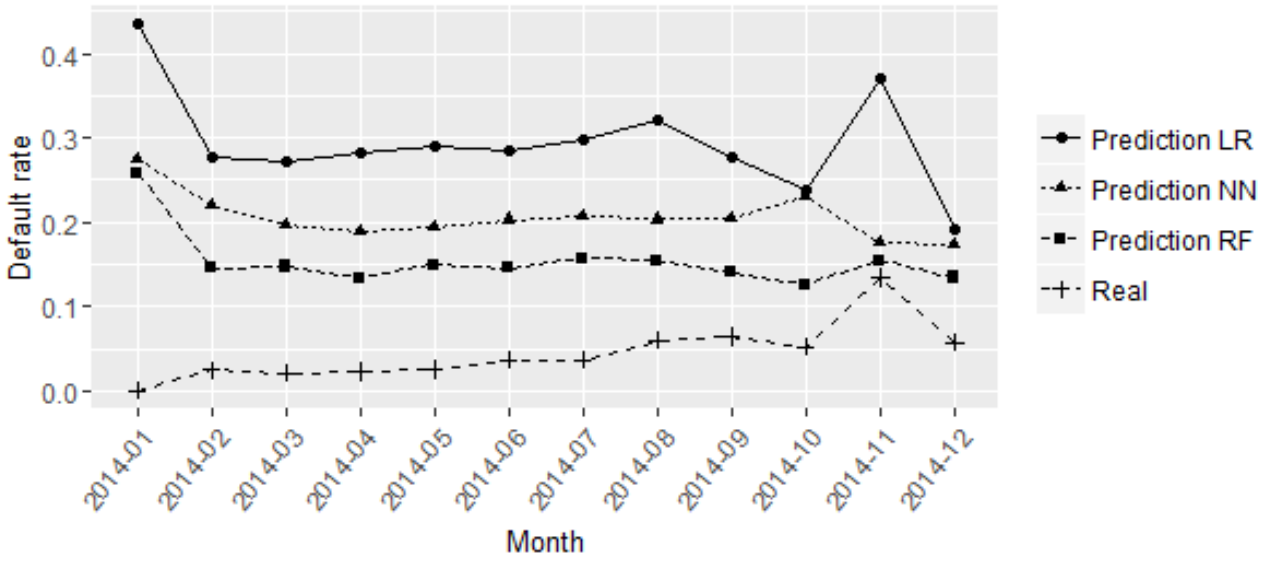


Figure 2: Default rate by month

## 6.3 Joint Analysis of Modelling Techniques, Explanatory Variables and Segmentation

In this section, we use a different approach to measure the impact of each group of variables on the performance of the model. Since the logistic regression is the most widely used approach in the literature, we focus our attention on this model for various segments of customers. We use the random forest approach to measure the importance of each explanatory variable because this model performs an implicit feature selection, using a subset of strong variables for the classification (Breiman, 2004). In particular, the Gini criterion (equivalent to the AUC) is used for measuring how well a split separates the samples in the two classes. The Gini criterion uses the Gini index, which is often used as a measure of income inequality. This index can be calculated as 1 minus twice the area between the Lorenz curve and the diagonal line representing perfect equality (values in the interval $[0, 1]$). In this way, a higher Gini index indicates greater discrimination between two classes.

The random forest model provides two measures of variable importance: the mean decrease Gini (MDG) and the mean decrease accuracy (MDA) (Calle & Urrea, 2011). The MDG is the sum of all the decreases in the Gini impurity due to a given variable, normalised by the number of trees. The MDA is the average accuracy of the predictor minus the decrease in the accuracy after the permutation of that predictor. We prefer using the MDG because its rankings are more robust than those generated using the MDA (Calle & Urrea, 2011).

Figure 3 displays the ranking of the 15 most important variables from the lowest to the highest MDG. For all the segments of customers, the crop type and the term type (payment frequency of the loan) are the most relevant, followed by various variables belonging to the credit and the behavioural groups. The main differences between the segments of clients, in relationship to the importance of the variables, are shown by the term type and the level of purchases. The term type is important for companies and persons, but not for holdings companies. In contrast, the level of purchases is relevant for companies and holding companies. Holding companies in particular show more significant variables in the agribusiness group, meaning that the economic conditions related to crops are more prevalent in this segment than in the others. This makes sense since the segment is oriented mostly to the SME, which tends to have higher variability and perceived risk (Maurer, 2014).

Table 12 shows that in logistic regression models the variables selected in most cases belong to the behavioural set, followed by the sociodemographic characteristics. Even if the agribusiness variables have been chosen in each segmentation, different segments are related to different agribusiness variables. For example, the crop type and the cost appear in the "Persons" segment, and the property distance shows up in all three segments. This hints at a diversity among the different segments that needs to be captured by different models.

Regarding neural networks, we used the variable importance method proposed by Garson

(1991). This method is based on connection weights to measure the relative importance of the explanatory variables in relation to the response variable. Table 13 presents the most important variables for neural networks models that consider all different types of variables. In this case, some sociodemographic and agribusiness variables related to incomes are the most important in all segments.

Another conclusion that can be drawn is that the risk brought by the liability amount is higher and relevant only for companies and holdings, since persons tend to have liabilities concentrated in a narrower and thus less significant range. However, the term is far more relevant for persons, which can be explained by the income uncertainty brought about by the extended time between sowing and harvesting/selling crops, which for small borrowers has a far more relevant impact on their solvency. This also impacts their liquidity in the face of unexpected events affecting profitability, which is not the case when they are compared to companies.

Table 12: The Most Significant Variables for the Logistic Regression Model by Segmentation of Customers

| Variable [a] | Variable group | All | Persons | Companies | Holdings | Total |
|---|---|---|---|---|---|---|
| Arrears3to6Months | bh | 1 | 1 | 1 | 1 | 4 |
| TimelyPay6Mto9M | bh | 1 | 1 | 1 | 1 | 4 |
| ArrearsLast3M | bh | 1 | 0 | 1 | 1 | 3 |
| TimelyPayLast3M | bh | 1 | 0 | 1 | 1 | 3 |
| TimelyPay3Mto6M | bh | 1 | 0 | 1 | 1 | 3 |
| RegionG1 | sd | 1 | 0 | 1 | 1 | 3 |
| LevelPurchases | sd | 1 | 1 | 0 | 1 | 3 |
| ArrearsIncreaseLast3M | bh | 1 | 1 | 1 | 0 | 3 |
| Tenure | ap | 1 | 0 | 1 | 1 | 3 |
| PropertyDistance | ag | 0 | 1 | 1 | 1 | 3 |
| CropTypeG2 | ag | 1 | 1 | 0 | 0 | 2 |
| Cost | ag | 1 | 1 | 0 | 0 | 2 |
| CropsNumber | ag | 1 | 0 | 1 | 0 | 2 |
| TimelyInstLast3M | bh | 0 | 1 | 0 | 0 | 1 |
| TotalBalance | ap | 1 | 0 | 0 | 0 | 1 |
| IncomeHectare | ag | 1 | 0 | 0 | 0 | 1 |
| Income | ag | 0 | 1 | 0 | 0 | 1 |

[a] The abbreviations *sd*, *ap*, *ag* and *bh* stand for sociodemographic variables, the credit variables, the agribusiness variables and the behavioural variables, respectively.

Table 13: The Most Important Variables for the Neural Network Model by Segmentation of Customers

| Variable | Variable group | All | Persons | Companies | Holding |
|---|---|---|---|---|---|
| LevelPurchases | sd | 0.045 | 0.047 | 0.046 | 0.040 |
| IncomeHectare | ag | 0.048 | 0.037 | 0.040 | 0.038 |
| RegionG1 | sd | 0.041 | 0.036 | 0.038 | 0.037 |
| CropTypeG2 | ag | - | 0.040 | 0.043 | 0.039 |
| CompanyTime | ap | - | 0.046 | 0.043 | - |
| PropertyDistance | ag | - | 0.046 | 0.040 | - |
| ArrearsLast3M | bh | 0.043 | - | - | 0.039 |
| RecentAccounts | ap | - | - | 0.034 | 0.042 |
| RatioArrearsAmountLast3M | bh | 0.053 | - | - | - |
| OfficeClientDist | ap | 0.047 | - | - | - |
| Income | ag | - | 0.042 | - | - |
| ArrearsAmount6Mto9M | bh | 0.042 | - | - | - |
| OfficeRegion | ap | - | 0.042 | - | - |
| Tenure | ap | 0.042 | - | - | - |
| ProductGroupNumber | ap | - | 0.041 | - | - |
| TermTypeG | ap | - | - | 0.041 | - |
| Cost | ag | - | 0.041 | - | - |
| TimelyPay3to6Months | bh | - | - | - | 0.040 |
| ArrearsAmount3Mto6M | bh | 0.039 | - | - | - |
| PayAmount6Mto9M | bh | 0.039 | - | - | - |
| ArrearsIncreaseAmount3Mto6M | bh | - | - | - | 0.039 |
| CurrencyG1 | ap | - | - | - | 0.037 |
| MinArrearsAmountLast3M | bh | - | - | - | 0.037 |
| PropertyLocationN | ag | - | - | 0.037 | - |
| ArrearsIncreaseAmountLast3M | bh | - | - | 0.036 | - |


Figure 3: Importance of variables

The abbreviations *sd*, *ap*, *ag* and *bh* stand for sociodemographic variables, the credit variables, the agribusiness variables and the behavioural variables, respectively.

## 6.4 Cost-Benefit Analysis

This section presents an analysis of the costs and benefits of using the model. These costs and benefits have been measured with a base scenario developed with Verbraken, Verbeke, and Baesens (2013) as the reference. The base scenario is the situation in which there is no classification model. In the case of credit scoring, this scenario occurs when all loans are granted; then this comparison ensures consistency when evaluating different credit scoring models (Verbraken et al., 2014).

We calculated the profit of using a model with the average classification profit per borrower (Verbraken et al., 2014)

$$P\left(t; b_1; c_0; c^*\right) = \left(b_1 - c^*\right) \pi_1 F_1\left(t\right) - \left(c_0 + c^*\right) \pi_0 F_0\left(t\right), \tag{4}$$

where functions $F_1(t)$ and $(F_0(t))$ are the cumulative density function of the scores of the cases and non-cases, respectively. The prior probabilities of classes 1 and 0 are $\pi_1$ and $\pi_0$ respectively. In relation to benefits and cost, $b_1$ is the benefit of correctly identifying a defaulter, $c_0 \geq 0$ is the cost of incorrectly classifying a good applicant as a defaulter, and $c^*$ is the cost of the action. We used the methodology developed by Bravo, Maldonado, and Weber (2013) to calculate each of

these parameters:

$b_1$ is calculated as the fraction of the loan amount that is lost after default.

$$b_1 = \frac{LGD \cdot EAD}{A}, \tag{5}$$

where $A$ is the principal, $LGD$ is the loss given default, and $EAD$ is the exposure at default.

$c_0$ is equal to the return on investment (ROI) of the loan, it is calculated by the cost of the funds and all operational costs.

$c^*$ is assumed $= 0$ because rejecting a customer does not generate costs.

The ROI of the company ($c_0$) is 0.05. Because the company does not have any sort of advance internal ratings-based approach (IRB), that is, its own internal estimates of risk components (Basel Committee on Banking Supervision, 2004), we set LGD equal to 1, defaulting to foundational IRB parameters (that mandate LGD = 1 for unsecured retail loans).

Results can be seen in table 14 and demonstrate that making use of a model leads to a utility greater than zero, that is, using a credit scoring tool is beneficial in economic terms. Specifically, the technique that has the biggest total profit is random forest (RF) and the best profit per loan (granted loans) is achieved by logistic regression (LR). In this sense, the technique can be selected according to the business objective, costs, and efforts of model development and implementation.

According to the results, using a credit score model is a good option in economic terms, regardless of the technique chosen. Scoring models can be used at different levels as a support tool in the lending decision, from a guide to classifying clients to the main method of evaluation, that is, by automatically accepting or rejecting clients according to their credit. First, an easy-to-interpret model could be better than a "black box" as a support tool in the loan decision. Logistic regression is the most interpretable technique of the three analysed, and this technique has a competitive performance compared to machine learning techniques. However, random forest is the most profitable option, despite the loss of interpretability. Therefore, the decision which model to use depends on the purpose and level of use of the credit score model.

Another important cost to consider is the cost of implementation. This cost could be divided into different aspects: computer infrastructure for the training model process and for future evaluations, expert knowledge for building the model, and training for the organisation in order to properly use the credit scoring. This cost increases as the complexity of the model increases.

A general recommendation is to start with an easy-to-interpret technique such as logistic regression and then migrate to a machine-learning technique. Random forest is a good option because it has good performance and the possibility to compute the importance of variables (mean decrease Gini and mean decrease accuracy), and its training process is less complex than other techniques, such as neural networks.

Table 14: Profit by Model

| Model | Profit (USD) | Granted loans | Profit per loan (USD) |
|---|---|---|---|
| LR | 230,067 | 1,995 | 115.32 |
| NN | 104,000 | 2,120 | 49.06 |
| RF | 234,930 | 2,554 | 91.99 |

# 7 Conclusions and Future Work

The credit risk assessment for the agricultural sector shows specific characteristics created by the uncertainty of successful crops and the lack of reliable information. Using data provided by a Chilean company, this study shows that the repayment behaviour characteristics and agribusiness variables are some of the most important aspects that contribute to causing farmers' repayment defaults. Among these groups, the most relevant variables are the days in arrears and the type of crop.

For the modelling technique, the random forest shows the best performance, followed by the widely used logistic regression model. There is a 6% gain between the best logistic regression and the best random forest, which suggests that the gain realised by exploiting more complex patterns is minor when compared to the gain in using better variable segments.

Concerning the segmentation of customers, the model estimated on the out-of-time sample of all the customers shows more stable results than those estimated on the segments of borrowers (persons, companies and holding companies). The main differences among these segments are given by the importance of the level of purchases and the agribusiness variables. The results clearly show how the patterns are structurally different among these segments, with variables that have distinct relevance. However, the predictive accuracy of a combined model is in line with a differentiated one, so a lender who does not desire to obtain the relevant information that comes from various models for each segment may choose to use only one model while the sophistication of the lender increases.

The previous result also leads to an interesting conclusion: given that we can draw a parallel between the size of the company and the reality of different countries (i.e. the purpose of the loans, type of borrowers, access to bank loans, among other aspects), we can see that in general, while the models need to be different for each reality (i.e. they require different variables), the statistical performance measures are similar. This is surprising, because one would expect that the greater data availability of larger, more sophisticated companies would lead to better capabilities to detect default, but the results seem to indicate that a dedicated lender who collects correct data will be able to detect this correctly across many segments (i.e. realities).

The main conclusion that can be drawn from this study is that a lender for agribusinesses does not face an extremely different scenario from that of a traditional lender. As long as the variables regarding the particular business are collected, and care is taken regarding which segments the

lender serves, it is possible to use existing credit-scoring technologies without further complexity. Doing so should provide an equivalent risk coverage to that of a lender serving a wider segment of the population and not facing any additional risks. Coverage in this segment should, then, also be equivalent to that of other groups of the population.

Regarding the limitations of the study, two can be pinpointed. First, the database most likely under-represents very-low-income and low-income countries. Even though there are low-income agribusinesses present in the data, they are sophisticated enough to gain access to formal suppliers, which occurs in low-to-middle-income countries and above (Reardon et al., 2009). Second, we are using traditional, structured databases, without any unstructured data (e.g. text, images, psychometric) that would require more sophisticated machine learning approaches. Perhaps, if these data were publicly available, potential gains shown here could be more significant.

Future work could include additional factors in the analysis, such as the impact of macroeconomic variables on the stability of the scoring models for the agribusiness sector. Another future development could be to improve the estimates of the agricultural incomes and costs to obtain estimates closer to actual values and to measure the impact of these estimates on the performance of the model.

# 8 Acknowledgements

The first and last authors acknowledge the support of Conicyt Fondecyt Initiation into Research Nº 11140264. The third author acknowledges that this research was undertaken, in part, thanks to funding from the Canada Research Chairs program.

# 9 Declaration of Interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

# References

Aruppillai, T., & Phillip, P. M. G. (2014). Farmers characteristics and its influencing on loans resettlement decision in Sri Lanka. *International Journal of Economics and Finance*, *6*(4), 110.

Baesens, B., Roesch, D., & Scheule, H. (2016). *Credit risk analytics: Measurement techniques, applications, and examples in SAS*. Hoboken, New Jersey: WILEY.

Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003, March). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science*, *49*(3), 312-329. doi: 10.1287/mnsc.49.3.312.12739

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*(6), 627-635.

Bandyopadhyay, A. (2008). Credit risk models for managing bank's agricultural loan portfolio. *ICFAI Journal of Financial Risk Management*, *5*(4), 86-102.

Barry, P. J., & Robison, L. J. (2001). Agricultural finance: Credit, credit constraints, and consequences. *Handbook of Agricultural Economics*, *1*, 513-571.

Basel Committee on Banking Supervision. (2004, June). International convergence of capital measurement and capital standards.

Becerra, W., Fiebig. (2004). *Agricultural production lending: A toolkit for loan officers and loan portfolio managers*. Eschborn, Germany: Pact Publications.

Bonazzi, G., & Iotti, M. (2014). Agricultural cooperative firms: Budgetary adjustments and analysis of credit access applying scoring systems. *Am. J. Appl. Sci*, *11*(7), 1181-1192.

Bravo, C., Maldonado, S., & Weber, R. (2013). Granting and managing loans for micro-entrepreneurs: New developments and practical experiences. *European Journal of Operational Research*, *227*(2), 358-366.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Breiman, L. (2004). *Consistency for a simple model of random forests* (Tech. Rep.). University of California at Berkeley.

Brida, J. G., Fasone, V., Scuderi, R., & Zapata-Aguirre, S. (2014). ClustOfVar and the segmentation of cruise passengers from mixed data: Some managerial implications. *Knowledge-Based Systems*, *70*, 128-136.

Calabrese, R., Osmetti, S. A., & Zanin, L. (2019, October). A joint scoring model for peer-to-peer and traditional lending: A bivariate model with copula dependence. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(4), 1163-1188. doi: 10.1111/rssa.12523

Calle, M. L., & Urrea, V. (2011). Letter to the editor: Stability of random forest importance measures. *Briefings in Bioinformatics*, *12*(1), 86-89.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 321-357.

Dressler, J. B., & Tauer, L. W. (2016). Estimating expected and unexpected losses for agricultural mortgage portfolios. *American Journal of Agricultural Economics*, *98*(5), 1470-1485. doi: doi:10.1093/ajae/aaw049

Durguner, S., & Katchova, A. L. (2007). Credit Scoring Models in Illinois by Farm Type: Hog, Dairy, Beef and Grain. *Urbana*, *51*, 61801.

EME. (2014). *Acceso a financiamiento en los emprendimientos [Access to financing in enterprises] (in spanish).*

Eyo, E., & Ofem, U. (2014). Analysis of creditworthiness and loan repayment among bank of agriculture loan beneficiaries (Poultry farmers) in Cross River State, Nigeria. *International Journal of Livestock Production*, *5*(9), 155-164.

Featherstone, A. M., Roessler, L. M., & Barry, P. J. (2006). Determining the probability of default and risk-rating class for loans in the seventh farm credit district portfolio. *Applied Economic Perspectives and Policy*, *28*(1), 4-23.

Fica, A. L. L., Casanova, M. A. A., & Mardones, J. G. (2018). Análisis de riesgo crediticio, propuesta del modelo credit scoring. *Revista de la Facultad de Ciencias Económicas: Investigación y Reflexión*, *26*(1), 181-207.

Gallagher, R. L. (2001). Characteristics of unsuccessful versus successful agribusiness loans. *Agricultural Finance Review*, *61*(1), 20-35.

Garson, D. (1991). Interpreting neural network connection weights. *AI Expert*, 47-51.

Gustafson, C. R., Pederson, G. D., & Gloy, B. A. (2005). Credit risk assessment. *Agricultural Finance Review*, *65*(2), 201-217.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

Hassoun, M. H. (1995). *Fundamentals of artificial neural networks.* USA: MIT Press.

Hazell, P. B. (1992). The appropriate role of agricultural insurance in developing countries. *Journal of International Development*, *4*(6), 567-581.

Henning, J. I., & Jordaan, H. (2016). Determinants of financial sustainability for farm credit applications A delphi study. *Sustainability*, *8*(1), 77.

Hosmer, D., & Lemeshow, H. (2000). *Applied logistic regression.* John Wiley & Sons.

Hou, J., Skees, J., & Wang, W. (2005). Potential of credit scoring in microfinance institutions in US (community venture corp. of kentucky taken as case study). In *Southern agricultural economics association 2005 annual meeting, february 5-9, 2005, little rock, arkansas.*

Jouault, A., & Featherstone, A. M. (2011). Determining the probability of default of agricultural loans in a French bank. *Journal of Applied Finance & Banking*, *1*(1), 1-30.

Katchova, A. L., & Barry, P. J. (2005). Credit risk models and agricultural lending. *American Journal of Agricultural Economics*, *87*(1), 194-205.

Kiers, H. A. (1991). Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, *56*(2), 197-212.

Klein, B., Meyer, R., Hannig, A., Burnett, J., & Fiebig, M. (2001). *Mejores Prácticas del Financiamiento Agrícola [Better Practices in Agricultural Finance] (In Spanish).* FAO.

Kullback, S. (1997). *Information theory and statistics.* USA: Dover Publications.

Lessmann, S., Seow, H.-V., Baesens, B., & Thomas, L. C. (2013). Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update. *European Journal of*

*Operational Research*, *247*(1), 124-136.

Limsombunchai, V., Gan, C., & Lee, M. (2005). An analysis of credit scoring for agricultural loans in Thailand. *American Journal of Applied Sciences*, *2*(8), 1198-1205.

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*(2), 145-151.

Madeira, C. (2019). Measuring the covariance risk of consumer debt portfolios. *Journal of Economic Dynamics and Control*, *104*, 21-38.

Mansfield, E. R., & Helms, B. P. (1982). Detecting multicollinearity. *The American Statistician*, *36*(3a), 158-160.

Maurer, K. (2014). Where is the Risk? Is Agricultural Banking Really More Difficult than Other Sectors? In D. Köhne (Ed.), *Finance for food* (p. 139-165). Springer.

Miller, B., Ellinger, & Lajili. (1993). Price and non-price management of agricultural credit risk. *Agricultural Finance Review*, *53*, 28-41.

Miller, L. H., & LaDue, E. L. (1988). *Credit assessment models for farm borrowers: A logit analysis* (Tech. Rep.). Cornell University, Department of Applied Economics and Management.

Novak, M., & LaDue, E. (1994). An analysis of multiperiod agricultural credit evaluation models for New York dairy farms. *Agricultural finance review (USA)*.

Novak, M. P., LaDue, E., et al. (1999). Application of recursive partitioning to agricultural credit scoring. *Journal of Agricultural and Applied Economics*, *31*, 109-122.

Odeh, O. O., Featherstone, A. M., Sanjoy, D., et al. (2006). Predicting credit default in an agricultural bank: Methods and issues. In *2006 annual meeting, february 5-8, 2006, orlando, florida.*

ODEPA. (2009). *Estudio de financiamiento agrícola.*

ODEPA. (2013). *Financiamiento agrícola: ¿Por qué surco debe seguir? [Agricultural financing: Which groove should it follow?](In Spanish).*

Onyenucheya, F., & Ukoha, O. (2007). Loan repayment and credit worthiness of farmers under the Nigerian Agricultural Cooperative and Rural Development Bank (NACRDB). *Agricultural Journal*, *2*(2), 265-270.

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, *74*, 26 - 39. doi: 10.1016/j.asoc.2018.10.004

Pelka, N., Musshoff, O., & Weber, R. (2015). Does weather matter? How rainfall affects credit risk in agricultural microfinance. *Agricultural Finance Review*, *75*(2), 194-212.

Probst, P., & Boulesteix, A.-L. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, *18*(181), 1-18.

Rambaldi, A. N., Zapata, H. O., Christy, R. D., et al. (1992). Selecting the" best" prediction model: An application to agricultural cooperatives. *Southern Journal of Agricultural Economics*, *24*,

163-163.

Reardon, T., Barrett, C. B., Berdegué, J. A., & Swinnen, J. F. (2009). Agrifood industry transformation and small farmers in developing countries. *World development*, *37*(11), 1717-1727.

Romani, G. A., Bravo, A., Aroca, P., Aguirre, N. A., Vega, P. L., & Carrazana, J. M. (2002). Modelos de clasificación y predicción de quiebra de empresas: Una aplicación a empresas chilenas. In *Forum empresarial* (Vol. 7, p. 1).

Römer, U., Römer, U., Musshoff, O., & Musshoff, O. (2017). Can agricultural credit scoring for microfinance institutions be implemented and improved by weather data? *Agricultural Finance Review*.

Savitha, B., Savitha, B., Kumar K, N., & Kumar K, N. (2016). Non-performance of financial contracts in agricultural lending: A case study from Karnataka, India. *Agricultural Finance Review*, *76*(3), 362-377.

Sherrick, B. J., Barry, P. J., & Ellinger, P. N. (2000). Valuation of credit risk in agricultural mortgages. *American Journal of Agricultural Economics*, *82*(1), 71-81.

Siddiqi, N. (2007). *Credit risk scorecards: Developing and implementing intelligent credit scoring* (Vol. 3). John Wiley & Sons.

Siddiqi, N. (2017). *Intelligent credit scoring: Building and implementing better credit risk scorecards*. John Wiley & Sons.

The World Economic Forum. (2017, September). The global competitiveness report 2017 – 2018.

Thomas, L., Crook, J., & Edelman, D. (2017). *Credit scoring and its applications, second edition*. Society for Industrial and Applied Mathematics. doi: 10.1137/1.9781611974560

Tiedemann, T., & Latacz-Lohmann, U. (2013). Production risk and technical efficiency in organic and conventional agriculture–the case of arable farms in Germany. *Journal of Agricultural Economics*, *64*(1), 73-96.

Turvey, C. G. (1991). Credit scoring for agricultural loans: A review with applications. *Agricultural finance review (USA)*, *51*, 43-54.

United Statistics Division. (2016). ISIC rev.4.

Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, *238*(2), 505-513.

Verbraken, T., Verbeke, W., & Baesens, B. (2013). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering*, *25*(5), 961-973.

Zech, L., & Pederson, G. (2003). Predictors of farm performance and repayment ability as factors for use in risk-rating models. *Agricultural Finance Review*, *63*(1), 41-54.

Zhang, T., & Ellinger, P. N. (2006). Credit Risk and Financial Performance Assessment of Illinois Farmers: A Comparison of Approaches with Farm Accounting Data. *Urbana*, *51*, 61801.

Ziari, H. A., Leatham, D. J., & Turvey, C. G. (1994). *Application of mathematical programming*

*techniques in credit scoring of agricultural loans* (Proceedings: 1994 Regional Committee NC-207, October 1994, Washington, DC No. 131475). Regional Research Committee NC-1014: Agricultural and Rural Finance Markets in Transition.

# 10 Appendix

Table 15: Summary statistics of continuous variables

| Variable | Group | Description |
|---|---|---|
| Cost | ag | Agricultural investment |
| CostHectare | ag | Agricultural activity cost per hectare |
| CostProperty | ag | Ratio between agricultural investment and no. of properties of the customer |
| CropsNumber | ag | No. of crop types |
| CropTypeG1 | ag | Crop type (first way to grouping) |
| CropTypeG2 | ag | Crop type (second way to grouping) |
| Income | ag | Estimated of total income of the agricultural holdings |
| IncomeHectare | ag | Income per hectare of the client's properties |
| IncomeProperty | ag | Ratio between the income and the number of the client's properties |
| ProductGroupG1 | ag | Main group of product according to purchases frequency (first way to grouping) |
| ProductGroupG2 | ag | Main group of product according to purchases frequency (second way to grouping) |
| PropertiesNumber | ag | No. of properties of the customer |
| PropertyDistance | ag | Avg. distance between each property and its nearest branch office |
| PropertyLocationN | ag | No. of property locations |
| PropertyRegionG | ag | Region associated with the property with the highest income of the customer |
| PropertyTypeG1 | ag | Property type of the property with the highest income |
| SurfaceProd | ag | Total plantation area in hectares |
| AccountBalance | ap | Amount owed by the customer in the account |
| AccountTime | ap | Account age in years |
| ActivityTime | ap | Elapsed time since the client is active in the company (years) |
| CompanyTime | ap | No. of years that the borrower has been a customer of the company |
| Currency | ap | Currency of the credit |
| Npurchases | ap | No. of purchases before the credit application |
| OfficeClientDist | ap | Distance between nearest branch office and customer's home |
| OfficeRegionG1 | ap | Branch office region of credit application (first way to grouping) |
| OfficeRegionG2 | ap | Branch office region of credit application (second way to grouping) |
| PreviousAccountsN | ap | No. of previous accounts in the company |
| ProductGroupNumber | ap | No. of product groups of the credit |
| ProductNumber | ap | No. of products purchased |
| RecentAccounts | ap | No. of new accounts (originated during last year) |
| Tenure | ap | If the client has a account in the company |
| TermType | ap | Term type according to payment frequency of the loan |
| TimeLastAccount | ap | Time elapsed since the most recent credit account application (years) |
| TimeLastMaturity | ap | Time elapsed since the most recent maturity of the customer's loans (months) |
| TimeLastUpdate | ap | Time elapsed since the most recent update in client's information (days) |
| TotalBalance | ap | Amount owed by the client in all his/her accounts |
| Arrears3to6Months | bh | Avg. of days in arrears in the last 3-6 months |
| Arrears6to9Months | bh | Avg. of days in arrears in the last 6-9 months |
| ArrearsAmount3Mto6M | bh | Avg. of arrears amount in the last 3-6 months |
| ArrearsAmount6Mto9M | bh | Avg. of arrears amount in the last 6-9 months |
| ArrearsAmountLast3M | bh | Avg. of arrears amount in the last 3 months |
| ArrearsDecrease3Mto6M | bh | No. of decreases in arrears amount in the last 3-6 months |
| ArrearsDecrease6Mto9M | bh | No. of decreases in arrears amount in the last 6-9 months |
| ArrearsDecreaseMto3M | bh | No. of decreases in arrears amount in the last 3 months |
| ArrearsIncrease3Mto6M | bh | No. of increases of the arrears amount in the last 3-6 months |
| ArrearsIncrease6Mto9M | bh | No. of increases of the arrears amount in the last 6-9 months |
| ArrearsIncreaseLast3M | bh | No. of increases of the arrears amount in the last 3 months |
| ArrearsLast3M | bh | Avg. of days in arrears in the last 3 months |
| MaxArrearsAmount3Mto6M | bh | Max. amount in arrears in the last 3-6 months |
| MaxArrearsAmount6Mto9M | bh | Max. amount in arrears in the last 6-9 months |
| MaxArrearsAmountLast3M | bh | Max. amount in arrears in the last 3 months |
| MaxArrearsDays3Mto6M | bh | Max. number of arrears days in the last 3-6 months |
| MaxArrearsDays6Mto9M | bh | Max. number of arrears days in the last 6-9 months |
| MaxArrearsDaysLast3M | bh | Max. number of arrears days in the last 3 months |
| MinArrearsAmount3Mto6M | bh | Min. amount in arrears in the last 3-6 months |
| MinArrearsAmount6Mto9M | bh | Min. amount in arrears in the last 6-9 months |
| MinArrearsAmountLast3M | bh | Min. amount in arrears in the last 3 months |
| MinArrearsdays3Mto6M | bh | Min. number of arrears days in the last 3-6 months |
| MinArrearsdays6Mto9M | bh | Min. number of arrears days in the last 6-9 months |
| MinArrearsdaysLast3M | bh | Min. number of arrears days in the last 3 months |
| NrPastDue3Mto6M | bh | No. of past due instalments in the last 3-6 months |
| NrPastDue6Mto9M | bh | No. of past due instalments in the last 6-9 months |
| NrPastDueLast3M | bh | No. of past due instalments in the last 3 months |
| NrTimely3Mto6M | bh | No. of timely instalments payments in the last 3-6 months |
| NrTimely6Mto9M | bh | No. of timely instalments payments in the last 6-9 months |
| NrTimelyLast3M | bh | No. of timely instalments payments in the last 3 months |
| PayAmount3Mto6M | bh | Avg. ratio between the payment and the instalment in the last 3-6 months |
| PayAmount6Mto9M | bh | Avg. ratio between the payment and the instalment in the last 6-9 months |
| PayAmountLast3M | bh | Avg. ratio between the payment and the instalment in the last 3 months |
| RatioArrearsAmount3M | bh | Ratio between the amount in arrears and its average value within the last 3 months |
| RatioArrearsAmount3Mto6M | bh | Ratio between the amount in arrears and its average value within the last 3-6 months |
| RatioArrearsAmount6Mto9M | bh | Ratio between the amount in arrears and its average value within the last 6-9 months |
| TimelyInst3Mto6M | bh | Avg. ratio between timely payment and instalment value in the last 3-6 months |
| TimelyInst6Mto9M | bh | Avg. ratio between timely payment and instalment value in the last 6-9 months |
| TimelyInstLast3M | bh | Avg. ratio between timely payment and instalment value in the last 3 months |
| TimelyPay3Mto6M | bh | Avg. ratio between timely payment and total payment in the last 3-6 months |
| TimelyPay6Mto9M | bh | Avg. ratio between timely payment and total payment in the last 6-9 months |
| TimelyPayLast3M | bh | Avg. ratio between timely payment and total payment in the last 3 months |
| ClientType | sd | Type of client (person, holding company, company) |
| EconomicActivityG1 | sd | Indicates the economic sector of the client (first way to grouping) |
| EconomicActivityG2 | sd | Indicates the economic sector of the client (second way to grouping) |
| LevelPurchases | sd | Classification of the customer according to his/her level of purchases |
| RegionG1 | sd | Geografic region of the client |

Table 16: Summary statistics of continuous variables

| Variable | Mean | Std.Dev | Min | Median | Max | N.Valid |
|---|---|---|---|---|---|---|
| Cost | 668.87 | 2,519.90 | - | 322.50 | 94,077.00 | 193,559 |
| CostHectare | 0.08 | 0.56 | - | 0.03 | 200.00 | 193,559 |
| CostProperty | 101.32 | 645.50 | - | 75.00 | 26,733.33 | 193,559 |
| CropsNumber | 3.07 | 1.89 | 1.00 | 3.00 | 11.00 | 193,559 |
| Income | 28,250.18 | 111,217.97 | - | 7,720.00 | 22,419,420.00 | 193,559 |
| IncomeHectare | 0.84 | 1.19 | - | 0.72 | 200.00 | 193,559 |
| IncomeProperty | 3,574.41 | 7,958.31 | - | 1,786.67 | 934,142.50 | 193,559 |
| PropertiesNumber | 7.84 | 9.92 | 1.00 | 4.00 | 99.00 | 193,559 |
| PropertyDistance | 33.94 | 89.69 | - | 16.71 | 2,196.55 | 193,559 |
| PropertyLocationN | 1.34 | 1.01 | 1.00 | 1.00 | 25.00 | 193,559 |
| SurfaceProd | 45.95 | 226.80 | 0.05 | 10.91 | 80,050.40 | 193,559 |
| AccountBalance | 6,569,036.87 | 16,959,771.75 | - | 2,170,840.00 | 358,668,412.00 | 193,559 |
| AccountTime | 1.02 | 0.91 | - | 1.00 | 4.00 | 193,559 |
| ActivityTime | 0.98 | 0.90 | - | 1.00 | 4.00 | 193,559 |
| CompanyTime | 9.11 | 5.53 | - | 9.00 | 23.00 | 193,559 |
| Npurchases | 42.59 | 55.24 | - | 24.00 | 472.00 | 193,559 |
| OfficeClientDist | 52.98 | 157.66 | - | 16.06 | 3,490.07 | 193,559 |
| PreviousAccountsN | 0.00 | 0.04 | - | - | 2.00 | 193,559 |
| ProductGroupNumber | 1.50 | 0.59 | 1.00 | 1.00 | 8.00 | 193,559 |
| ProductNumber | 2.20 | 1.72 | 1.00 | 2.00 | 20.00 | 193,559 |
| RecentAccounts | 0.06 | 0.25 | - | - | 2.00 | 193,559 |
| TimeLastAccount | 375.29 | 355.38 | - | 346.00 | 1,501.00 | 193,559 |
| TimeLastMaturity | 122.88 | 140.17 | - | 55.00 | 1,222.00 | 193,559 |
| TimeLastUpdate | 352.80 | 277.58 | - | 287.00 | 1,501.00 | 193,559 |
| TotalBalance | 6,570,458.33 | 16,959,629.77 | - | 2,171,795.00 | 358,668,412.00 | 193,559 |
| Arrears3to6Months | 9.26 | 16.82 | - | 2.21 | 348.00 | 193,559 |
| Arrears6to9Months | 7.64 | 15.68 | - | - | 454.00 | 193,559 |
| ArrearsAmount3Mto6M | 127,570.58 | 232,200.95 | - | 29,812.00 | 9,377,517.00 | 193,559 |
| ArrearsAmount6Mto9M | 104,976.99 | 218,671.39 | - | - | 9,377,517.00 | 193,559 |
| ArrearsAmountLast3M | 148,077.07 | 252,587.86 | - | 65,851.00 | 12,143,950.00 | 193,559 |
| ArrearsDecrease3Mto6M | 3.77 | 8.04 | - | 1.00 | 127.00 | 193,559 |
| ArrearsDecrease6Mto9M | 3.10 | 7.27 | - | - | 125.00 | 193,559 |
| ArrearsDecreaseMto3M | 4.50 | 8.52 | - | 1.00 | 125.00 | 193,559 |
| ArrearsIncrease3Mto6M | 4.14 | 9.29 | - | - | 156.00 | 193,559 |
| ArrearsIncrease6Mto9M | 3.40 | 8.48 | - | - | 151.00 | 193,559 |
| ArrearsIncreaseLast3M | 4.84 | 9.64 | - | - | 156.00 | 193,559 |
| ArrearsLast3M | 10.91 | 17.51 | - | 5.29 | 474.00 | 193,559 |
| MaxArrearsAmount3Mto6M | 577,534.05 | 1,059,134.76 | - | 117,848.00 | 13,915,766.00 | 193,559 |
| MaxArrearsAmount6Mto9M | 469,869.27 | 971,677.80 | - | - | 14,161,000.00 | 193,559 |
| MaxArrearsAmountLast3M | 668,031.12 | 1,102,486.08 | - | 293,187.00 | 14,161,000.00 | 193,559 |
| MaxArrearsDays3Mto6M | 19.40 | 29.63 | - | 8.00 | 1,036.00 | 193,559 |
| MaxArrearsDays6Mto9M | 15.87 | 27.70 | - | - | 454.00 | 193,559 |
| MaxArrearsDaysLast3M | 22.94 | 31.16 | - | 16.00 | 1,036.00 | 193,559 |
| MinArrearsAmount3Mto6M | 12,394.89 | 90,283.69 | - | - | 9,377,517.00 | 193,559 |
| MinArrearsAmount6Mto9M | 10,224.39 | 88,908.07 | - | - | 9,377,517.00 | 193,559 |
| MinArrearsAmountLast3M | 14,131.28 | 122,021.10 | - | - | 12,143,950.00 | 193,559 |
| MinArrearsdays3Mto6M | 2.74 | 10.46 | - | - | 348.00 | 193,559 |
| MinArrearsdays6Mto9M | 2.30 | 9.68 | - | - | 454.00 | 193,559 |
| MinArrearsdaysLast3M | 3.17 | 10.98 | - | - | 474.00 | 193,559 |
| NrPastDue3Mto6M | 8.27 | 16.18 | - | 2.00 | 166.00 | 193,559 |
| NrPastDue6Mto9M | 6.84 | 15.16 | - | - | 166.00 | 193,559 |
| NrPastDueLast3M | 9.96 | 17.43 | - | 3.00 | 175.00 | 193,559 |
| NrTimely3Mto6M | 3.74 | 7.90 | - | - | 113.00 | 193,559 |
| NrTimely6Mto9M | 3.05 | 6.89 | - | - | 99.00 | 193,559 |
| NrTimelyLast3M | 4.56 | 8.87 | - | 1.00 | 152.00 | 193,559 |
| PayAmount3Mto6M | 0.67 | 0.47 | - | 1.00 | 1.00 | 193,559 |
| PayAmount6Mto9M | 0.55 | 0.50 | - | 1.00 | 1.00 | 193,559 |
| PayAmountLast3M | 0.80 | 0.40 | - | 1.00 | 1.00 | 193,559 |
| RatioArrearsAmount3M | 0.67 | 1.60 | - | - | 64.56 | 193,559 |
| RatioArrearsAmount3Mto6M | 1.69 | 61.20 | - | - | 8,995.78 | 193,559 |
| RatioArrearsAmount6Mto9M | 2.71 | 104.98 | - | - | 12,689.47 | 193,559 |
| TimelyInst3Mto6M | 0.26 | 0.35 | - | - | 1.00 | 193,559 |
| TimelyInst6Mto9M | 0.22 | 0.34 | - | - | 1.00 | 193,559 |
| TimelyInstLast3M | 0.31 | 0.36 | - | 0.17 | 1.00 | 193,559 |
| TimelyPay3Mto6M | 283.62 | 8,656.95 | - | - | 984,843.00 | 193,559 |
| TimelyPay6Mto9M | 524.19 | 31,241.10 | - | - | 3,599,348.00 | 193,559 |
| TimelyPayLast3M | 785.96 | 33,601.72 | - | 0.17 | 3,599,348.00 | 193,559 |

Table 17: Summary statistics of categorical variables

| Variable | Values | Frequency | Frequency (%) |
|---|---|---|---|
| RegionG1 | North | 13,887 | 7.17% |
| | Metropolitan | 7,423 | 3.84% |
| | South | 2,324 | 1.20% |
| | Valparaíso | 27,717 | 14.32% |
| | O'Higgins | 90,682 | 46.85% |
| | Maule | 30,338 | 15.67% |
| | Ñuble | 21,188 | 10.95% |
| EconomicActivityG1 | Agribusiness | 187,294 | 96.76% |
| | Others | 6265 | 3.24% |
| EconomicActivityG2 | Agribusiness group 0: fruit, livestock, transport, services, poultry. | 9,158 | 4.73% |
| | Agribusiness group 1: forestry, plants, exports. | 178,371 | 92.15% |
| | Others | 6,030 | 3.12% |
| LevelPurchases | A: purchases exceeding 73,500 USD | 64,489 | 33.32% |
| | B: purchases between 22,050 and 73,500 USD | 59,357 | 30.67% |
| | C: minor purchases to 22,050 USD. | 69,713 | 36.02% |
| ClientType | Companies | 69,692 | 36.01% |
| | Holding Companies | 64,295 | 33.22% |
| | Person | 59,572 | 30.78% |
| PropertyRegionG | North | 16,096 | 8.32% |
| | South | 1,989 | 1.03% |
| | Valparaíso and Metropolitan | 16,658 | 8.61% |
| | O'Higgins | 99,948 | 51.64% |
| | Maule | 35,501 | 18.34% |
| | Ñuble | 22,714 | 11.73% |
| | Los Lagos | 653 | 0.34% |
| CropTypeG1 | Annual crops | 4,972 | 2.57% |
| | Fruits | 58,418 | 30.18% |
| | Vegetable | 20,667 | 10.68% |
| | Corn | 28,341 | 14.64% |
| | Others | 19,352 | 10.00% |
| | Wheat | 8,023 | 4.14% |
| | Grapes | 28,001 | 14.47% |
| | Vineyard | 25,785 | 13.32% |
| CropTypeG2 | Annual crops | 4,972 | 2.57% |
| | Fruits group 0: Almond, blueberries, cherry, custard apple, plums, clementine, peach, raspberry, strawberry, lemon, mandarino, apple cantaloupe, blackberry, pears. | 30,158 | 15.58% |
| | Fruits group 1: Citrus, kiwis, watermelon and others. | 28,687 | 14.82% |
| | Vegetables group 0: Chicory, garlic, artichoke, asparagus, lettuce, avocado, potato, cabbage, tomato, carrot and others. | 8,270 | 4.27% |
| | Vegetables group 1: Onion, pepper, pumpkin. | 12,648 | 6.53% |
| | Corn | 28,341 | 14.64% |
| | Others | 18,586 | 9.60% |
| | Wheat | 8,023 | 4.14% |
| | Grapes | 28,001 | 14.47% |
| | Vineyard group 0: export grapevine, Pisco grapevine, vinifera, Cabernet vineyard, Merlot vineyard. | 11,106 | 5.74% |
| | Vineyard group 1: Others grapevines. | 14,767 | 7.63% |
| PropertyTypeG1 | Rent | 47,985 | 24.79% |
| | Commodatum | 546 | 0.28% |
| | Own | 137,830 | 71.21% |
| | Usufruct | 7,198 | 3.72% |
| Tenure | 0: If the client doesn't have an account in the company | 169,765 | 87.71% |
| | 1: If the client has a account in the company | 23,794 | 12.29% |
| OfficeRegionG1 | North zone | 23,030 | 11.90% |
| | Central zone | 62,298 | 32.19% |
| | Central South zone 1 | 82,381 | 42.56% |
| | Central South zone 2 | 23,657 | 12.22% |
| | South zone | 2,193 | 1.13% |
| OfficeRegionG2 | North zone | 23,030 | 11.90% |
| | Central zone 1 | 31,318 | 16.18% |
| | Central zone 2 | 30,980 | 16.01% |
| | Central South zone 1a | 37,245 | 19.24% |
| | Central South 1b | 45,136 | 23.32% |
| | Central South zone 2 | 23,657 | 12.22% |
| | South zone | 2,193 | 1.13% |
| Currency | CLP | 173,737 | 89.76% |
| | USD | 19,822 | 10.24% |
| TermType | 0-30 days | 70,103 | 36.22% |
| | 31-60 days | 68,548 | 35.41% |
| | 61-90 days | 8,817 | 4.56% |
| | 91-120 days | 17,964 | 9.28% |
| | More than 121 days | 28,127 | 14.53% |
| ProductGroupG1 | Product group 1 | 21,461 | 11.09% |
| | Product group 2 | 54,922 | 28.37% |
| | Product group 3 | 53,879 | 27.84% |
| | Product group 4 | 63,297 | 32.70% |
| ProductGroupG2 | Product group 1 | 21,461 | 11.09% |
| | Product group 2 | 54,922 | 28.37% |
| | Product group 3-4 | 117,176 | 60.54% |