UNIVERSIDAD DE TALCA

FACULTAD DE INGENIERÍA

ANALYSIS OF THE IMPACT OF BEHAVIORAL AND SECTOR-SPECIFIC VARIABLES IN CREDIT RISK MEASUREMENT FOR THE AGRIBUSINESS

por

Daniela Antonieta Lazo Toledo

Tesis para optar al grado de

Magíster en Gestión de Operaciones

Profesor Supervisor Cristián Bravo Román

Septiembre, 2016

CONSTANCIA

La Dirección del Sistema de Bibliotecas a través de su encargado Biblioteca Campus Curicó certifica que el autor del siguiente trabajo de titulación ha firmado su autorización para la reproducción en forma total o parcial e ilimitada del mismo.

Curicó, 2019

Magíster en Gestión de Operaciones

FACULTAD DE INGENIERÍA

ACTA EXAMEN DEFENSA DE TESIS GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES

TITULO: ANALYSIS OF THE IMPACT OF BEHAVIORAL AND SECTOR-SPECIFIC VARIABLES IN CREDIT RISK MEASUREMENT FOR THE AGRIBUSINESS

SEÑORITA: DANIELA ANTONIETA LAZO TOLEDO

Nº MATRÍCULA: 2015801002

AÑO DE INGRESO: 2015

AÑO DE TÉRMINO: 2016

	NOTA DE BORRADOR (60%)	NOTA DE PRESENTACIÓN (40%)
CALIFICACIONES **Observación General:** La escala de notas es de 1,0 a 7,0, siendo 4,0 la nota mínima de aprobación.	7,0	7,0

CALIFICACIONES:

Nota Final: 7.0

siete, cero

Nota Final (en letras)

COMISIÓN EXAMINADORA

Raffaella Calabrese

DRA. RAFFAELLA CALABRESE
(PROFESOR EXTERNO)

DR. ALEJANDRO RODRÍGUEZ
(FACULTAD DE INGENIERÍA)

DR. CRISTIÁN BRAVO ROMÁN
(PROFESOR SUPERVISOR)

DRA. MARCELA GONZÁLEZ ARAYA
PRESIDENTE DE LA COMISIÓN

UNIVERSIDAD DE TALCA
Programa Magister en Gestión Operaciones
DIRECTOR
CAMPUS CURICO

CURICÓ, 28 DE SEPTIEMBRE DE 2016

Resumen Ejecutivo

El sector agrícola contribuye con un 6,1% del PIB mundial total. China, India y EE.UU. son los mayores participantes con un 34,58% del PIB agrícola total (The World Factbook, 2015). Los empresarios del sector agrícola tienen características específicas relacionadas con los préstamos, debido a los ciclos agrícolas y los riesgos morales. Dado que no es posible conocer el rendimiento y la consecuente ganancia de los cultivos, es importante que con la información disponible se trate de predecir el comportamiento del cliente al momento del vencimiento del préstamo.

Este documento presenta un estudio del impacto de los principales factores relevantes para este segmento, relacionados con el riesgo de crédito, proporcionando una visión sobre el enfoque que los prestamistas del agronegocio deben tomar para proporcionar mejores servicios financieros al sector.

Los datos utilizados provienen de una empresa chilena que otorga crédito a agricultores para el suministro de insumos, y corresponden a 6.658 clientes que fueron aprobados entre enero de 2007 y diciembre de 2013.

El análisis del riesgo de crédito en el agronegocio se realiza considerando tres factores simultáneamente: el tipo de cliente (personas, empresas y holdings), la técnica de modelización (*Random Forests*, Redes Neuronales y Regresión Logística), y la información disponible (sociodemográfica, de comportamiento de pago, de agronegocio y del crédito).

Los resultados muestran que los patrones son estructuralmente diferentes entre los segmentos de clientes, con variables que tienen una gran relevancia; Sin embargo, la precisión predictiva de un modelo combinado está en línea con un modelo diferenciado. Por otro lado, incluir las variables de comportamiento aumenta el AUC entre 5% y 20%, en el caso de las variables de agronegocio el incremento es entre 5%-10%. *Random Forests* fue el mejor modelo en general, sin embargo la Regresión Logística tiene un buen desempeño y ofrece a los prestamistas agrícolas una manera fácil para medir los riesgos de crédito, teniendo en cuenta variables especializadas en el proceso de modelado.

Como trabajo futuro se podría incluir factores adicionales en el análisis, tales como el impacto de las variables macroeconómicas sobre la estabilidad de los modelos de puntuación para el sector agrícola. Otro desarrollo futuro podría ser mejorar las estimaciones de los ingresos agrícolas y los costos para obtener estimaciones más cercanas a los valores reales y medir el impacto de estas estimaciones en el desempeño del modelo.

Palabras Claves: Puntuación de crédito, Agronegocio, Comportamiento de pago.

Extended Abstract

The agricultural sector contributes with a 6.1% of total world GDP. China, India and the US are the best contributors with a 34.58% of the total agricultural GDP (The World Factbook, 2015). Entrepreneurs of the agricultural sector have specific features related to lending, because the agricultural cycles and the moral risks. Since it is not possible to know the performance and subsequent gain of the crops, it is important that with the information available try to predict customer behavior at maturity.

This paper presents a study of the impact of the main factors relevant to this segment providing insights on the focus that agribusiness-oriented lenders have to take in order to provide better financial services to the sector.

The data used comes from a Chilean company that provides credit to farmers for the supply of inputs, it contains 6,658 customers who were approved between January 2007 and December 2013.

The analysis of credit risk in agribusiness is performed considering three different factors simultaneously: company size (persons, companies and holding companies), modeling technique (Random Forests, Neural Networks and Logistic Regression), and available information (socio-demographic, repayment behavior, agribusiness-specific and credit-related).

The results clearly show how the patterns are structurally different among the customer segments, with variables that have distinctly relevance; however, the predictive accuracy of a combined model is in line with a differentiated one. On another hand, including behavioral variables increases AUC by between 5%-20%, in case of agribusiness variables the increment is by between 5%-10%. Random Forests was the best model overall, nevertheless Logistic Regression has good performance and deliver an easy way for agricultural lenders to measure credit risks, considering specialized variables in modeling process.

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.

Key Words: Credit Scoring, Agribusiness, Repayment Behavior.

Contents

Index of Tables

Index of Figures

CHAPTER 1: INTRODUCTION

1 Introduction

1.1 Background and problem

The agricultural sector contributes with a 6.1% of total world GDP. China, India and the US are the best contributors with a 34.58% of the total agricultural GDP (The World Factbook, 2015).

Entrepreneurs of the agricultural sector have specific features related to lending. The information on borrowers who have low incomes is difficult to obtain (Becerra, 2004). The agricultural production is inherently a risky business (Hazell, 1992). The agriculture has long production cycles: this production cycles typically reflect in loans with seasonal terms (Barry, 2001).

Since it is not possible to know the performance and subsequent gain of the crops, it is important that with the information available (generally less information than desired by banks) try to predict customer behavior at maturity.

1.2 Objetives of the study

The aim of the study consists in establish which are the most important factors in default of farmers and provide a methodology and recommendations to agricultural lenders. The specific objectives of the research are:

- Analysis of the impact of creation of specialized variables in the credit risk of agricultural lending, considering the available information of the farmers by input suppliers (less information than banks).
- We focus in the creation of the agribusiness and repayment behavior variables and measure the contribution of these in the prediction of default of farmers.

1.3 Contribution

To the best of our knowledge, this study is the first to analyze the most appropriate methodology to model credit risk in the agribusiness considering three different factors simultaneously: company size, modeling technique, and available information.

We analyze the type of classification techniques in order to determine if is it worth it to develop different techniques to classical model (Logistic Regression) and we also evaluated if it is convenient to use different models for each segment of clients.

1.4 Methodology

The data used comes from a Chilean company that provides credit to farmers for the supply of inputs, it contains 6,658 customers who were approved between January 2007 and December 2013. We use a sample equivalent to 161,613 credit sales.

The experimental design of the study consists of a full factorial experimental setup in order to assess the effects of three different factors on the performance of prediction of default of the farmers.

- Type of variables: socio-demographic, repayment behavior, agribusiness-specific and credit-related. Sociodemographic variables regarding the characteristics of the client such as age, geographic region, goods and disposable income. Credit related variables consist of the basic information of the conditions of the loan, such us term, number of years that the borrowers has been a customer of the company, payment type (semesterly, monthly, annually), etc. Agribusiness variables consist of attributes in relation to crops, productivity, farm yield and farm size. Behavioral variables are related to the credit history of the clients.

- Type of clients: i.e. if a customer represents a person, an enterprise or a holding company. We have used the classification provided by the lender. We aim to explore the impact of the borrower in the probability of default of the farmers.

- Model Techniques: Random Forests, Neural Networks and Logistic Regression. Random Forests (Breiman, 2001) are a powerful and robust alternative to predict default, due to their great ability to detect complex patterns, they have been shown to be the most powerful ensemble model to predict credit scorecards (Lessmann et al., 2013). To capture nonlinear relationships we use Neural Networks, a powerful but difficult to interpret model (Hassoun, 1995). Logistic Regression, is selected for its simplicity that does not sacrifice too much its discriminatory power, given that this model performs very well for credit scoring (Baesens et al., 2003).

1.5 Main results and conclusions

According to the results, the repayment behaviour variables and agribusinesses are important in explaining the default of farmers. Behavioural variables are the most important, but a mix of variables is required. In general, including behavioural variables increases AUC by between 5%-20%, in case of agribusiness variables the increment is by between 5%-10%.

In relation to the models, Random Forests is the best model overall, using all variable sets. The classical model (Logistic Regression) has a good result that is competitive with machine learning models.

The customer results by segment indicate that exists an increase of 3% in the out-of-sample of using a segmented model, but there is a decrease of 3% in the out-of-time sample. Regarding to the different segments of clients the model including all customers has more stable results. The results show that different segments do have significantly different risk behaviour, enough to merit creating different scoring models for each group rather than just including a segmentation variable.

Entrepreneurs in the agricultural sector have specific characteristics associated with the loans: it is important that with the information available try to predict customer behavior at maturity. Logistic Regression models deliver an easy way for interconnected systems to measure credit risks, but they have to be tailored to their customers. On other hand, the repayment behavior variables and agribusinesses are important in explaining the default of farmers.

1.6 Description of structure of the document

The organization of the document is as follow: Section 2.1 is an introduction that presents the main literature of agribusiness credit scoring and the contribution of the paper. Section 2.2 presents an agricultural credit scoring literature review, Section 0 describes the financing sources available for farmers in developing countries and shows the main financing sources in Chile. Section 0 presents data considerations while Section 2.5 describes the credit scoring methodology to be applied. Section 2.6 shows the main results. Finally, Section 00 shows the conclusions of the study and future work proposals.

CHAPTER 2: ANALYSIS OF THE IMPACT OF BEHAVIORAL AND SECTOR-SPECIFIC VARIABLES IN CREDIT RISK MEASUREMENT FOR THE AGRIBUSINESS

2 Analysis of the impact of behavioral and sector-specific variables in credit risk measurement for the agribusiness

Análisis del impacto de las variables de comportamiento y sectoriales en la medición del riesgo de crédito para el agronegocio

Abstract

This work provides insight on the focus that agribusiness-oriented lenders have to take in order to provide better financial services to the sector. Entrepreneurs in the agricultural sector represent a large portion of companies worldwide. This paper presents a study of the impact of the main factors relevant to this segment. The data used comes from a Chilean company that provides credit to farmers for the supply of inputs, it contains 6,658 customers who were approved between January 2007 and December 2013. The analysis of credit risk in agribusiness is performed considering three different factors simultaneously: company size (persons, companies and holding companies), modeling technique (Random Forests, Neural Networks and Logistic Regression), and available information (socio-demographic, repayment behavior, agribusiness-specific and credit-related). The results show that including behavioral variables increases AUC by between 5%-20%, in case of agribusiness variables the increment is by between 5%-10%. Random Forests was the best model overall, nevertheless Logistic Regression has good performance and deliver an easy way for agricultural lenders to measure credit risks, considering specialized variables in modeling process.

Key Words: Credit Scoring, Agribusiness, Repayment Behavior.

2.1 Introduction

Entrepreneurs of the agricultural sector have specific features related to lending. The most important risks in agricultural lending are related to moral hazard and associated to factors with the agricultural sector and production (Becerra, 2004).

On relation to moral hazard, farmers have more knowledge about their production risk than credit institutions. In this point, the main problem is that the information on borrowers who have low incomes is difficult to obtain (Becerra, 2004).

The agricultural production is inherently a risky business (Hazell, 1992). The agricultural production risks are related to pest, disease and weather, because of these risks the production could not have the expected returns, which reflects in default by borrowers. The agriculture has long production cycles: in this time the market prices may change from what has been projected (Becerra, 2004). Further, these production cycles typically reflect in loans with seasonal terms (quarterly, semiannual, annual, etc.)

(Barry, 2001).

This work focuses on entrepreneurs of the agricultural sector. To define this sector we use The International Standard Industrial Classification of All Economic Activities (ISIC), the international reference classification of productive activities, that include forestry, hunting and fishing, crops and livestock production (United Statistics Division, 2016).This sector contributes 6.1% of total world GDP. China, India and the US are the best contributors with a 34.58% of the total agricultural GDP (The World Factbook, 2015).

When studying variables for measuring the credit risk for farmers, the typical approach is using financial ratios (Miller and LaDue, 1988; Rambaldi et al., 1992; Novak et al., 1999; Jouault and Featherstone, 2011). Some authors used different variables from financial ratios, for example Gallagher (2001) used financial characteristics, and non-financial characteristics, as manager and lender experience, in a logit model and found that there is an improvement in accuracy by adding non-financial characteristics. Hou (2001) included demographic statistics and their business and loan information, using a logistic regression analysis determined that most of the significant variables correspond to the latter category. Limsombunchai et al. (2005) determines the lending decision is a function of borrower characteristics, credit risk proxies, relationship indicators, and dummy variables about agribusiness and loan information, the results of the logistic credit scoring model show the significance of credit risks proxies. Aruppillai and Phillip (2014) showed that socioeconomic characteristics improve the efficiency of the lending decision.

Other relevant factor in credit risk is segmentation, that is the division the clients in groups according to some criterion. In some cases, using several scorecards (by a segmentation of costumers), provides better risk differentiation than using just one scorecard on everyone (Siddiqi, 2007). In credit scoring for the agribusiness there are different segmentations, for example current and not current loans (Ziari et al., 1994), loan size (Miller and LaDue, 1988), type of activity or product (Bandyopadhyay et al., 2007), loan type (Bandyopadhyay et al., 2007), etc.

To the best of our knowledge, our study is the first to analyze the most appropriate methodology to model credit risk in the agribusiness considering three different factors simultaneously: company size, modeling technique, and available information (socio-demographic, repayment behavior, agribusiness-specific and credit-related). Additionally, we analyze the characteristics of the measurement of credit risk in farmers in developing countries, using the example of Chile, considering the forms of financing in developing countries, especially the interconnected systems.

The data used to construct the model comes from a Chilean company that provides credit to farmers for the supply of inputs, besides providing support services. Funding sources which also serve as input suppliers (with multiple offices close to their customers) have an advantage in customer

closeness and knowledge of different agricultural specialties (ODEPA, 2013). Furthermore, most the customers have different size of agricultural crops and varied incomes, and this allows for a comparison between different types of clients; therefore, the exploration of this data has significant potential in determining the relevant factors in this segment.

This paper presents an analysis of the impact of creation of specialized variables in the credit risk of agricultural lending, considering the available information of the farmers by input suppliers (less information than banks). We focus in the creation of the agribusiness and repayment behavior variables and measure the contribution of these in the prediction of default of farmers. Also we analyze the type of classification techniques in order to determine if it is worth it to develop different techniques to classical model (Logistic Regression) and the company size. Finally, the aim of the study consists in establish which are the most important factors in default of farmers and provide a methodology and recommendations to agricultural lenders.

2.2 Measuring Credit Risk in the Agribusiness

Credit risk is a primary source of risk to financial institutions and the holdings of capital, the main responses to this risk are loan loss allowances and equity assets (Barry and Robison, 2001).

Information about past financial performance is the dominant signal agricultural borrowers can provide to distinguish their credit risks Miller and Lajili (1993).However, data limitations are a major impediment in assessing farm financial performance (Zhang and Ellinger, 2006). Regarding to small farmers, their business scale, geographic remoteness, informal accounting practices, and their business risks and financial risks, imply high information needs to allow lenders to adequately manage credit risks (Barry and Robison, 2001).

Several studies have examined the credit risk in agribusiness. A number of these studies use portfolio credit risk management models used 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 et al. (2006) used credit scoring techniques to rate a portfolio of loans. Sherrick et al. (2000) and Dressler and Tauer (2016) developed credit-risk valuation models to measure the credit risks to estimate expected and unexpected losses. Other studies assess credit risks of individual loans through credit scoring models (Miller and LaDue, 1988; Turvey, 1991; Novak et al., 1999; Hou, 2001). The literature of credit scoring is very limited compared to the portfolio analysis literature (Thomas et al., 2002). Our study focuses in credit scoring models because the aim of this work is assess the risk in lending to individual customers and assisting in loan approval decisions to agricultural lenders, thus the literature review of the following paragraphs

is focused in credit scoring models.

Regarding the credit models that have been used for assessment the credit risk in the agricultural sector, those include Logistic Regression (Miller and LaDue, 1988; Rambaldi et al., 1992; Novak et al., 1999; Hou, 2001; Limsombunchai et al., 2005; Durguner and Katchova, 2007), Discriminant Analysis (Rambaldi et al., 1992; Ziari et al., 1994), and machine learning techniques such as Decision Trees and Neural Networks (Novak et al., 1999; Limsombunchai et al.,2005).

The Logistic Regression model is the classic and most widely used due to its simplicity and explanatory power (Thomas, 2000). Machine learning models are complex, but they have advantages: for example, decision trees allow to establish the importance of the variables and Neural Networks can establish nonlinear relationships between them.

On relation to parametric and no-parametric models Ziari et al. (1994) found that either mathematical programming techniques or statistical models performed equally well, and that mixed integer-programming models perform better than parametric models. An advantage of non-parametric models is that these can fit several distribution functions. Furthermore, when the data sample is small or the data is contaminated, non-parametric models like Neural Networks may behave better (Gustafson et al., 2005).

The Logistic Regression is the technique more applied in agricultural credit scoring (See Table 1), however some studies present a comparison of different classification techniques. Turvey (1991) used a data of the Canada's Farm Credit Corporation to compare the performance of four credit scoring models (Linear Probability Model, Discriminate Analysis, Logit and Probit) and found that similar classification accuracies (between 71.5% and 67.1%) for these models. Odeh et al. (2006) compared Logistic Regression, Artificial Neural Networks and Adaptive Neuro-Fuzzy Inference (ANFI) system to predict default using data from Farm Credit System, identifying slight differences in prediction accuracies, ANFI was better than the other methods in sensitivity and specificity measures.

The types of variables used in the literature on credit scoring for farmers are mainly referred to financial ratios such as liquidity, profitability and leverage (Jouault and Featherstone, 2011; Ziari et al., 1994; Durguner and Katchova, 2007), farmer characteristics (educational level, age, goods etc.) (Limsombunchai et al., 2005), farm characteristics (types of crops, farm size) (Miller and LaDue, 1988; Novak et al., 1999; Limsombunchai et al., 2005; Onyenucheya and Ukoha, 2007), credit features and credit history (Jouault and Featherstone, 2011; Hou, 2001; Aruppillai and Phillip, 2014; Eyo and Ofem, 2014).

Turvey (1991) stresses the importance of inclusion of qualitative and quantitative attributes in the credit scoring models. Gallagher (2001) indicates that a prediction model without non-financial

variables could have model misspecification. Zech and Pederson (2003) identified the debt-to-asset ratio as a major predictor of repayment ability. Zech and Pederson (2003) also argued that the total asset turnover ratio and family living expenses are strong predictors of farm financial performance.

Table 1 presents a summary of the evaluation work credit for farmers in terms of model types, variables and the country applied for the loan. There are only a few analyses of all factors affecting the failure of farmers, most studies analyze different types of models or variables, but do not include the analysis of the factors simultaneously.

Limsombunchai et al. (2005) and Eyo and Ofem (2014) analyze two different models and types of variables, but do not take into account the size of the company and behavioral variables. This paper presents an analysis of the impact of the creation of specialized variables (agribusiness and repayment behavior), the type of classification techniques and company size simultaneously. This analysis is performed in order to determine which are the most important factors in default of farmers and recommendations to agricultural lenders on relation to credit risk.

Table 1: Credit scoring models for farmers.The models that were applied were: Logistic Regression (LR), Discriminant Analysis (DA), variations of Discriminant Analysis (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 utilization, profitability, assets, operational efficiency.	USA
Ziari et al. (1994)	DA, FLDA, LDA	Financial ratios	USA
Novak et al. (1999)	RPA, LR	Debt-to-asset ratio, current ratio.	USA
Hou (2001)	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)	BLR	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

2.3 Financing farmers in developing countries

According to FAO (2001), the types of rural lenders that can be found in developing countries are the following:

- Formal lenders: banks, agricultural development, 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 that make it to present high and complex risk profiles (ODEPA, 2009). Other sources of funding, especially interconnected systems (suppliers of agricultural inputs/crop buyers) “have an advantage in relation to customer closeness and knowledge of different fields, attributes that are valued beyond the rate interest charge” (ODEPA, 2013).

Referring to Chile, 17.9% of farmers use some form of credit for financing 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%), the second most important source of financing corresponds to suppliers of agricultural inputs (11.6%).

Using data from farmers seeking loans in credit interconnected systems can allow to determine the relevant factors in this segment, referred to their repayment behavior. This is due to the knowledge of the agricultural area and the closeness that these institutions have with their customers.

Table 2: Amount of farmers loans in Chile. Original data from (ODEPA, 2013).

Source	Amount (mlm of USD)	Percentage
Indap	69.81	1.10%
Input suppliers	711.16	11.60%
Agriculture contract	68.90	1.10%
Comodity exchange	53.87	0.90%
Foreign investement	39.51	0.60%
Credit unions	11.35	0.20%
Factoring	4.17	0.10%
Subtotal	958.77	15.60%
Banks	5,192.60	84.40%
Total	6,151.37	100.00%

2.4 Data

This section describes data base creation process for the scorecard. This database includes a set of predictor variables and a target variable. Section 2.4.1 describes the main considerations of the data preparation as data acquisition and sampling. Section 2.4.2 presents a description of the list of variables and Section 2.4.3 shows the variable selection and transformation process.

2.4.1 Data preparation

We have used data provided by a Chilean company that grant loans to farmers for the supply of inputs, besides providing support services. The data was anonymized to protect customer confidentiality and identity. It contains 6,658 customers who were approved between January 2007 and December 2013. The data includes a subset of their application characteristics and full subsequent repayment behavior up to December 2014.

We use a sample equivalent to 161,613 credit sales. The person scorecard has 48,875 cases, company scorecard has 58,443 cases and the holding scorecard has 54,295 cases. The default for the sample of all cases is 2.55%, the rates by segment are 2.56%, 2.48%, 2.64%, for persons, companies and holdings respectively.

In the Scorecards development, opened accounts in a time frame are used for predicting the performance of future accounts. This time frame is denominated “Performance Window”. We will use a typical performance windows for behavioral scorecards: 12 months (Siddiqi, 2007). Over this period we also construct the target variable.

The target variable corresponds to the following Good/Bad definition Anderson, 2007): the bad state corresponds to default and the default definition is according to international regulations (Basel Committee on Banking Supervision, 2004). This definition raises that the obligor is in default if past due more than 90 days on any material credit obligation.

In addition, it is necessary prepare the database to ensure data quality and reliability. At this point those variables which with lower variability or have more than 30% of missing values are deleted.

2.4.2 Database description

The list of variables has been generated by the first elimination: concentrated values (variables with low variability, which more than 95% of the cases are of a category), missing values (variables with more than 30% of missing values were eliminated) and the creation of derived variables by functions (minimum, maximum or mean), difference or ratio between variables. Regarding to sociodemographic we use region of the client residence, economic activity, level of purchases and type of client.

Agribusiness variables are related with incomes, cost, crop types and the ratios and transformations with this type of characteristics. Credit variables are attributes of the loan and the history of the customer in the company (for example time of tenure of the client). Respect to behavioral variables, three time windows were used: the last 3 months, the period from the last 3-6 months and the period from the last 6-9 months. We use the maximum, minimun, average, count of increments, count of the decrements and ratios with variables related to repayment behavior as arrears amount and days in arrears. We use these statistics with the aim of to measure the behavior of the clients in an aggregate value, since values of behavioral variables change over the performance window. In total we have 4 sociodemographic variables, 11 agribusiness-related variables, 16 credit variables and 42 behavioral variables.

2.4.3 Variable selection and transformation

The variable selection process had two stages. For measure the independence of the variables with the target variable, univariate tests have been applied: χ^2 and the KS test for categorical and continuous variables respectively. Through this univariate analysis we eliminated the variables that don't have relation with the target variable. Furthermore, we use clusters of variables for reduce the dimensions of the data, specifically we utilizes the ClustOfVar (Brida et al., 2014) algorithm, this algorithm applies kmeans clustering to categorical and continuous variables using as a center a synthetic variable calculated by Principal Component Analysis (Kiers, 1991).

In case of Logistic Regression models, the subset of variables was obtained after applying a correlation filter, multicollinearity through of variance inflation factors (Mansfield and Helms, 1982)and finally by Stepwise Selection. The models have variables with a significance level $\Pr(>|z|)$ less than 0.05 (we remove the variables that have a significance level higher than 0.05 in each iteration of Stepwise Selection).

Using the variable selection we reduce from 73 variables to 30 for the data set of the all costumers, 33 for the person dataset, 32 for the companies one and 29 for the holding dataset.

Respect to the variable transformation, we recode the features by weight of evidence (WOE) as a measure of the strength of the attributes to predict the target variable, this steps is common in credit scoring models, and allows for normalizing the dataset to the amount of information that each variable provides (Siddiqi, 2007). The WOE is calculated as follows:

$$\mathrm{WOE} = \ln\left(\frac{DistrGood}{DistrBad}\right)$$

Equation 1: WOE

Where $DistrGood$ and $DistrBad$ are the proportion of cases of the attribute that belongs to the

good and bad class respectively over there the total cases of the class.

For this transformation we first discretize the continuous variables using classification trees and regrouped the attributes of categorical variables so that all of categories has ad cases and at least 5% of the total cases.

2.5 Experimental design

The experimental design of the study consists of a full factorial experimental setup in order to assess the effects of three different factors on the performance of prediction of default of the farmers.

The first factor of the experimental design concerns the classification techniques and has three possible levels, one per technique that is applied. The classification techniques that we will use are explained in Section 2.5.1. The second factor is the types of variables, consisting of four possible levels (credit, behavioral, sociodemographic and agribusiness) and combinations thereof, these are presented in Section 2.5.2. The third factor represents the type of clients in three possible levels (single company, holding, natural person) which are described in Section 2.5.3. The objective of the study is to contrast the different levels and combinations of these three factors in order to determinate the impact of the agribusiness variables and other factors.

The models (constructed by the combinations of the factors) are compared in terms of the area under the receiver operator characteristic curve (AUC), a common methodology to contrast different models (Lobo et al., 2008).

Finally, we select the best model according to the previous measures. In addition, we perform the scorecard validation, in order to check that the model does not have overfitting through cross validation.

2.5.1 Classification techniques

We selected three different techniques to construct the credit scorecard, a linear and two nonlinear models are selected. The first model, Logistic Regression, is selected for its simplicity, which does not sacrifice too much its discriminatory power, given that this model performs very well for credit scoring (Baesens et al., 2003).

To capture nonlinear relationships we use Neural Networks, a powerful but difficult to interpret model (Hassoun, 1995).

Random Forests (Breiman, 2001) are a powerful and robust alternative to predict default, due to

their great ability to detect complex patterns. They have been shown to be the most powerful ensemble model to predict credit scorecards (Lessmann et al., 2013). Random Forests are classifiers which combine decision trees such that each of them uses a separate sample of the data. The sets of training data are generated with the bootstrapping method that generates new data sets of the same size using sampling with replacement, this allows reducing the variance and makes a more robust model.

2.5.2 Variable Sets

We construct four variable subsets. The first subset consists of sociodemographic variables regarding the characteristics of the client such as age, geographic region, goods and disposable income. Credit related variables consist of the basic information of the conditions of the loan, such us term, number of years that the borrowers has been a customer of the company, payment type (semesterly, monthly, annually), etc. Agribusiness variables consist of attributes in relation to crops, productivity, farm yield and farm size. Behavioral variables are related to the credit history of the clients.

Behavioral variables were constructed over three time windows (3, 6, 9 months). We measure the following variables over each of the time windows: amount in arrears and days in arrears. We calculate the average, minimum, maximum, number of increases and number of decreases for each measure and time frame. An example a behavioral variable would then be “average arrears amount within the last n months”.

Formally, let $\boldsymbol{x}$ be the set of all variables, let x_{ag} be the subset of agribusiness variables, x_{sd} the subset of sociodemographic variables, x_{ap} be the subset of credit variables and x_{bh} the subset of behavioral variables. The probability of default $P\,(y = 1|\boldsymbol{x})$ in the model that incorporates the three subsets of variables the result of estimating a function $f(\cdot)$ such that:

$$f\left(x_{ag}, x_{ap}, x_{sd}, x_{bh}\right) = P\,(y = 1|\boldsymbol{x})$$

Equation 2: Probability of default

2.5.3 Type of borrowers

The third factor to analyze is the type of client, i.e. if a customer represents a person, an enterprise or a holding company. We have used the classification provided by the lender. We aim to explore the impact of the borrower in the probability of default of the farmers.

Natural persons refers to customers who apply for credit individually and not are associated or belong to any company. The remaining categories, enterprises and holding companies, are clients who represents a company and a business organization that controls a number of companies, respectively.

We will now build models using all combinations of variables and factors. In total 15x3x3 (135) models are built and benchmarked using AUC, as we present in Section 2.6.

2.6 Results

In this section we show the results of the application of the methodology. First, we describe the initial characteristic analysis recommends by Siddiqi (2007) implies measuring the strength of each variable in relation to the target variable and determining whether the features have a logical relation with the variable to predict. Then, we present the model results and an analysis of importance of variables.

2.6.1 Initial characteristic analysis

After the recoding of the variables by WOE, we analyze the variables in relation to the ability to distinguish between good and bad cases. We determine whether the trend of a variable is logical (or not) using business knowledge, verified by members of the company providing credit. Thus, if the variable does not have a logical trend, its categories are regrouped. In this stage, we regrouped the attributes by business logic or similar WOE.

We calculate the information value (Kullback, 1968) of each variable to measure the total strength of the variable. The results for the model of all costumers can be seen in the Table Table 3. The groups “sd”, “ap”, “ag” and “bh” are equivalent to sociodemographic, credit, agribusiness and behavioral groups respectively, the information value and the classification of strength are according to Siddiqi (2007). The variables with high values of information value corresponds to the groups behavioral and agribusiness, in particular: arrears_day_1_90, arrears_days_91_180, timely_payment_1_90, Croptype_g2 and timely_installment_1_90. These variables indicate that the default of customers is related more with your recent payment behavior (previous 3-6 months) and the type of crop.

Table 3: Information Value

Variable	Group	Information value	Strength
arrears_days_1_90	Behavioral	0.56	Strong
arrears_days_91_180	Behavioral	0.44	Strong
timely_payment_1_90	Behavioral	0.30	Strong
CropType_g2	Agribusiness	0.29	Strong
timely_installment_1_90	Behavioral	0.29	Strong
TotalBalance	Credit	0.24	Strong
timely_payment_91_180	Behavioral	0.21	Strong
timely_payment_181_270	Behavioral	0.21	Strong
timely_installment_181_270	Behavioral	0.20	Strong
n_timely_1_90	Behavioral	0.20	Medium
Region_g1	Sociodemographic	0.18	Medium
Cost	Agribusiness	0.17	Medium
LevelPurchases	Sociodemographic	0.13	Medium
IncomeHectare	Agribusiness	0.13	Medium
CostProperty	Agribusiness	0.13	Medium
arrears_amount_1_180	Behavioral	0.12	Medium
CropsNumber	Agribusiness	0.12	Medium
Income	Agribusiness	0.11	Medium
arrears_amount_181_270	Behavioral	0.11	Medium
inc_arrears_amount_l_90	Behavioral	0.10	Medium
CostHectare	Agribusiness	0.10	Medium
n_past_due_1_90	Behavioral	0.10	Medium
inc_arrears_amount_91_180	Behavioral	0.09	Weak
Tenure	Credit	0.09	Weak
PropertyLocationN	Agribusiness	0.08	Weak
PropertyDistance	Agribusiness	0.07	Weak
PreviousPurchasesN	Credit	0.06	Weak
n_past_due_181_170	Behavioral	0.04	Weak
dec_arrears_amount_91_180	Behavioral	0.02	Weak
dec_arrears_amount_181_270	Behavioral	0.02	Weak
TimeLastMaturity	Credit	0.01	Unpredictive

2.6.2 Model results

Because the data sets are imbalanced respect to the classes of the target variable, we applied SMOTE (synthetic minority over-sampling technique), a technique that combines oversampling and undersampling (Chawla et al., 2002), for generating balanced data sets.

For validation of the models we create two data sets: An out-of-sample set equivalent to the 30% of the sample and an out-of-time sample with credit sales made January from 2014. The model parameters were adjusted by grid search (in each combination of variables and segment of clients) for Neural Networks and Random Forests, using 20% of the training sample to adjust those parameters.

The AUC results are reported in the Tables Table *4*, Table *5*,Table *6* and Table *7*. Each model has a combination of types of variables. Note that in most of cases the models include the different types of variables, which aims at the conclusion that information from all groups has to be included in order to maximize the model performance. If we constrain each model to include only one type of variables, then the models that perform best are the ones with behavioral variables, followed by agribusiness-related variables.

In Tables Table *8*,Table *9*,Table *10* and Table *11* we present a normalized values of AUC for the models of all customers (models without segmentation), persons, companies and holdings respectively. These values were calculated dividing by the maximum value of the out-of-time AUC for each segment. In general, including behavioral variables increases AUC by between 5%-20%, in case of agribusiness variables the increment is by between 5%-10%. In the models of all customers, if we see the Logistic Regression results, aggregating behavioral variables has bigger impact than the application of other models as Random forests. Something similar happens in the other segments of clients.

Table 12 shows the best results (AUC) for each segment for the out-of-sample and the out-of-time sample. The customer results by segment indicate that exists an increase of 3% in the out-of-sample of using a segmented model, but there is a decrease of 3% in the out-of-time sample, this indicates that using a one-size-fits-all model delivers a more stable result. However, in the Logistic Regression results of out-of-time and out-of-sample are quite similar (see Tables Table *4*, Table *5*, Table *6* and Table *7*)

On relation to the model relevance Random Forests is the best model overall, using all variable sets. Logistic Regression also has good results, Neural Networks have the worst result in the out-of-time sample and this can be seen in the Figure 1, in the ROC curve of the out-of-time sample for the model of all costumers.

Table 4: AUC Results of All costumers

Model Variables	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	0.810	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**

Table 5: AUC Results of Persons

Model Variables	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

Table 6: AUC Results of Companies

Model Variables	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**

Table 7: AUC Results of Holding Companies

Model Variables	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	0.654	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**

Table 8: Normalized AUC of All customers

Model Variables	Logistic Regression	Neural Networks	Random Forests
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

Table 9: Normalized AUC of Persons

Model Variables	Logistic Regression	Neural Networks	Random Forests
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

Table 10: Normalized AUC of Companies

Model Variables	Logistic Regression	Neural Networks	Random Forests
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

Table 11: Normalized AUC of Holding Companies

Model Variables	Logistic Regression	Neural Networks	Random Forests
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

Table 12: Best Results by segment

AUC	Out of sample	Out of time
Persons	0.938	0.854
Companies	0.939	0.870
Holdings	0.944	0.863
All	0.917	0.879

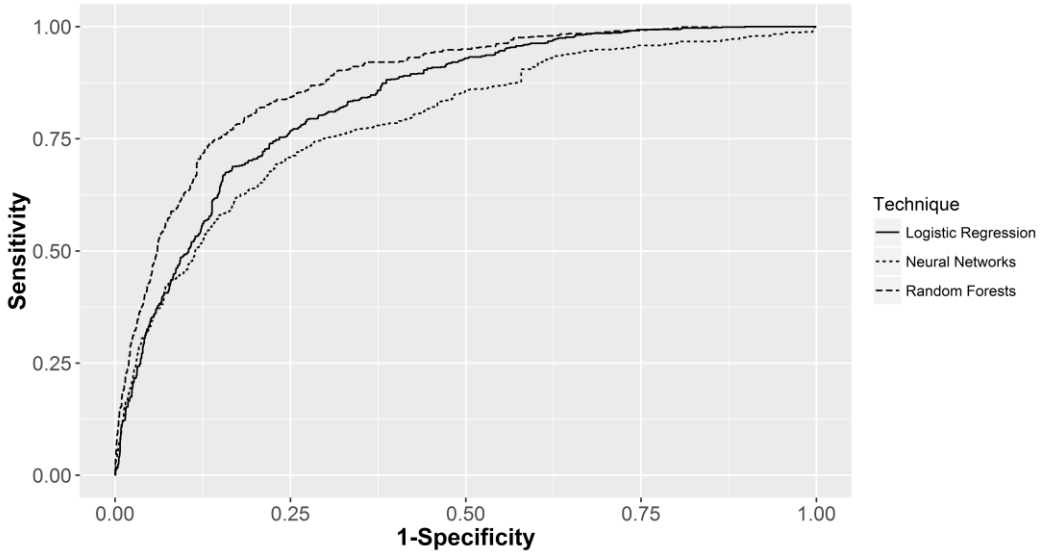

Figure 1: ROC curve of the all customers dataset (out-of-time sample)

2.6.3 Importance of variables

In this stage we present different approaches for asses the relative importance of the variables in the models.

First, in relation to Logistic Regression, we present the subset of variables of the models obtained for each segment.

On other hand, we use the Random forests models to measure the importance of variables. Random Forests performs an implicit feature selection, using a subset of strong variables for the classification (Breiman, 2004), the Gini criterion is used for measuring how well a split separates the samples in the two classes. Random Forests provides two variable importance measures: mean decrease Gini (MDG) and mean decrease accuracy (MDA) (Calle and Urrea, 2011). The MDG, is the sum of all decreases in Gini impurity, due to a given variable, normalized by the number of trees. The Mean decrease accuracy (MDA) is the average across of the accuracy for the predictor minus the decrease in accuracy after permutation of the predictor. We use the MDG because the rankings based on the MDG are more robust than MDA (Calle and Urrea, 2011).

In Table 13 can be seen a summary of the variables of the Logistic Regression models and their presence in the different models for the client segments. The variables with major presence in the models are of behavioral type, also the sociodemographic variables are presents in the models of all segments. The agribusiness variables are presents in all of models, nevertheless different segments have different agribusiness variables. For example, the crop type and cost appear in the persons model, and property distance appear in the models of three segments.

Regarding to the importance of variables in Random Forests, in Figure 2, the 15 most important variables by segment can be seen (higher values of mean decrease Gini are associated with an important variable). For all segments the variables of the crop type and term type are important, the next important variables are of credit or behavioral groups.

The main differences of the segments of clients on relation to the importance of the variables in the models are in the variable of the term type and level of purchases. Term type is important for companies and persons, but not for holdings companies. On relation to the level of purchases, this variable is more important for companies and holdings companies. Further, companies segment has more important variables of the type agribusiness variables. These differences show that in case of enterprises the repayment behavior is affected by level of purchases (persons generally has less values of purchases) and in case of smaller costumers the term type, which could be explained by the long periods in which generated uncertainty regarding the solvency of farmers to the end of the crops, where persons generally have less liquidity to face unexpected events affecting profitability compared to companies.

Table 13: Presence of variables in Logistic Regression models

Variable	Variable type	All	Persons	Companies	Holdings	Total
arrears_days_91_180	bh	1	1	1	1	4
timely_payment_181_270	bh	1	1	1	1	4
arrears_days_1_90	bh	1	0	1	1	3
timely_payment_1_90	bh	1	0	1	1	3
timely_payment_91_180	bh	1	0	1	1	3
Region_g1	sd	1	0	1	1	3
LevelPurchases	sd	1	1	0	1	3
inc_arrears_amount_1_90	bh	1	1	1	0	3
Tenure	ap	1	0	1	1	3
PropertyDistance	ag	0	1	1	1	3
CropType_g2	ag	1	1	0	0	2
Cost	ag	1	1	0	0	2
CropsNumber	ag	1	0	1	0	2
timely_installment_1 _90	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

Figure 2: Importance of variables

2.7 Conclusions and future work

Entrepreneurs in the agricultural sector have specific characteristics associated with the loans. Since it is not possible to know the performance and subsequent gain of the crops, it is important that with the information available (generally less information than desired by banks) try to predict customer behavior at maturity.

According to the results, the repayment behavior variables and agribusinesses are important in explaining the default of farmers. Of these the most important variables are referred to the days in arrears and crop type.

In relation to the models, the classical model (Logistic Regression) has a good result that is competitive with machine learning models, this by performing a selection of variables to avoid multicollinearity and selecting only significant variables in the model. Logistic Regression models deliver an easy way for interconnected systems to measure credit risks, but they have to be tailored to their customers considering specialized variables in modeling process.

Regarding to the different segments of clients (persons, companies and holdings) the model including all costumers has more stable results. The main differences of these segments in importance of variables are related to level of purchases and agribusiness variables, then, include this variables in

credit scoring models represent an advantage in the prediction of default.

Future work might be including more factors in the experiments. Macroeconomic variables such as exchange rate or GDP could provide more stable predictions. Also the development of a method to improve the estimate of agricultural incomes and cost, to get closer to the actual value, might also improve the results obtained.

Moreover, given the importance of crop type variable, it would be interest to consider (for larger lenders with sufficient loans) models segmented by this criterion.

2.8 Acknowledgements

The first and last author acknowledge the support of Conicyt Fondecyt Initiation into Research 11140264.

2.9 References

Anderson, Raymond. *The Credit Scoring Toolkit.* Oxford University Press, 2007.

Aruppillai, Thayaparan, and Paulina Mary Godwin Phillip. "Farmers Characteristics and Its Influencing on Loans Resettlement Decision in Sri Lanka." *International Journal of Economics and Finance* 6 (2014): 110-117.

Baesens, Bart, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, and Jan Vanthienen. "Benchmarking state-of-the-art classification algorithms for credit scoring." *Journal of the Operational Research Society* (Nature Publishing Group) 54 (2003): 627-635.

Bandyopadhyay, Arindam, and others. "Credit risk models for managing bank's agricultural loan portfolio." *ICFAI Journal of Financial Risk Management, Dec2008* (Citeseer) 5 (2007): 86-102.

Banking Supervision, Basel Committee. "International Convergence of Capital Measurement and Capital Standards." June 2004.

Barry, Peter J. "Modern capital management by financial institutions: implications for agricultural lenders." *Agricultural Finance Review* (MCB UP Ltd) 61 (2001): 103-122.

Barry, Peter J., and Lindon J. Robison. "Agricultural finance: Credit, credit constraints, and consequences." *Handbook of agricultural economics* (Elsevier) 1 (2001): 513-571.

Becerra, Fiebig, Wisniwski. "Agricultural Production Lending: A Toolkit for Loan Officers and Loan Portfolio Managers." Eschborn, 2004.

Breiman, Leo. "Consistency for a simple model of random forests." Tech. rep., 2004.

Breiman, Leo. "Random forests." *Machine learning* (Springer) 45 (2001): 5-32.

Brida, Juan Gabriel, Vincenzo Fasone, Raffaele Scuderi, and Sandra Zapata-Aguirre. "ClustOfVar and the segmentation of cruise passengers from mixed data: Some managerial implications." *Knowledge-Based Systems* (Elsevier) 70 (2014): 128-136.

Bryant, Kay. "ALEES: an agricultural loan evaluation expert system." *Expert systems with applications* (Elsevier) 21 (2001): 75-85.

Calle, M. Luz, and Víctor Urrea. "Letter to the editor: stability of random Forests importance measures." *Briefings in bioinformatics* (Oxford Univ Press) 12 (2011): 86-89.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research*, 2002: 321-357.

Division, United Statistics. "ISIC Rev.4." 2016.

Dressler, Jonathan B., and Loren W. Tauer. "Estimating Expected and Unexpected Losses for Agricultural Mortgage Portfolios." *American Journal of Agricultural Economics* (Oxford University Press), 98(2016):1470-1485

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

EME. "Acceso a Financiamiento en los Emprendimientos." [Access to Financing in Enterprises] (In Spanish). 2014.

Eyo, E. O., and U. I. Ofem. "Analysis of creditworthiness and loan repayment among bank of agriculture loan beneficiaries (Poultry farmers) in Cross River State, Nigeria." *International Journal of Livestock Production*, 2014.

FAO. "Mejores prácticas del financiamiento agrícola [Agricultural funding: which course to take?] (In Spanish)." 2001.

Featherstone, Allen M., Laura M. Roessler, and Peter J. Barry. "Determining the probability of default and risk-rating class for loans in the seventh farm credit district portfolio." *Applied Economic Perspectives and Policy* (Oxford University Press) 28 (2006): 4-23.

Gallagher, Richard L. "Characteristics of unsuccessful versus successful agribusiness loans." *Agricultural Finance Review* (MCB UP Ltd) 61 (2001): 20-35.

Gustafson, Cole R., Glenn D. Pederson, and Brent A. Gloy. "Credit risk assessment." *Agricultural Finance Review* (Emerald Group Publishing Limited) 65 (2005): 201-217.

Hassoun, Mohamad H. *Fundamentals of artificial neural networks.* MIT press, 1995.

Hazell, Peter B. R. "The appropriate role of agricultural insurance in developing countries." *Journal of International Development* (Wiley Online Library) 4 (1992): 567-581.

Henning, Johannes I. F., and Henry Jordaan. "Determinants of Financial Sustainability for Farm Credit Applications—A Delphi Study." *Sustainability* (Multidisciplinary Digital Publishing Institute) 8 (2016): 77.

Hou, Mr Jiang. "Potential of Credit Scoring in Microfinance Institution in US." *Agricultural Finance Review*, 2001.

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

Katchova, Ani L., and Peter J. Barry. "Credit risk models and agricultural lending." *American Journal of Agricultural Economics* (Oxford University Press) 87 (2005): 194-205.

Kiers, Henk A. L. "Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables." *Psychometrika* (Springer) 56 (1991): 197-212.

Kullback, Solomon. *Information theory and statistics.* Courier Corporation, 1968.

Lessmann, Stefan, Hsin-Vonn Seow, Bart Baesens, and Lyn C. Thomas. "Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update." *Credit Research Centre, Conference Archive.* 2013.

Limsombunchai, Visit, Christopher Gan, and Minsoo Lee. "An analysis of credit scoring for agricultural loans in Thailand." *American Journal of Applied Sciences* 2 (2005): 1198-1205.

Lobo, Jorge M., Alberto Jiménez-Valverde, and Raimundo Real. "AUC: a misleading measure of the performance of predictive distribution models." *Global ecology and Biogeography* (Wiley Online Library) 17 (2008): 145-151.

Mansfield, Edward R., and Billy P. Helms. "Detecting multicollinearity." *The American Statistician* (Taylor & Francis) 36 (1982): 158-160.

Miller, Ellinger, Barry, and Lajili. "Price and non-price management of agricultural credit risk." *Agricultural Finance Review*, 1993.

Miller, Lynn H., and Eddy L. LaDue. "Credit Assessment Models for Farm Borrowers: A Logit Analysis." Tech. rep., Cornell University, Department of Applied Economics and Management, 1988.

Novak, Michael P., Eddy LaDue, and others. "Application of recursive partitioning to agricultural credit scoring." *Journal of Agricultural and Applied Economics* (Southern Agricultural Economics Association) 31 (1999): 109-122.

Odeh, Oluwarotimi O., Allen M. Featherstone, Das Sanjoy, and others. "Predicting Credit Default in an Agricultural Bank: Methods and Issues." *Southern Agricultural Economics Associa-tion Annual Meeting, Orlando, FL.* 2006.

ODEPA. "Estudio de Financiamiento Agrícola." 2009.

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

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

Pederson, Glenn, and Nicholas Sakaimbo. "Default and loss given default in agriculture." *Agricultural*

Finance Review (Emerald Group Publishing Limited) 71 (2011): 148-161.

Rambaldi, Alicia N., Hector O. Zapata, Ralph D. Christy, and others. "Selecting The" Best" Prediction Model: An Application To Agricultural Cooperatives." *Southern Journal of Agricultural Economics* (Southern Agricultural Economics Association) 24 (1992): 163-163.

Sherrick, Bruce J., Peter J. Barry, and Paul N. Ellinger. "Valuation of credit risk in agricultural mortgages." *American Journal of Agricultural Economics* (Oxford University Press) 82 (2000): 71-81.

Siddiqi, Naeem. *Credit risk scorecards: developing and implementing intelligent credit scoring.* Vol. 3. John Wiley & Sons, 2007.

The World Factbook. "GDP - Composition, by sector of origin" 2015.

Thomas, Lyn C. "A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers." *International journal of forecasting* (Elsevier) 16 (2000): 149-172.

Thomas, Lyn C., David B. Edelman, and Jonathan N. Crook. *Credit scoring and its applications.* Siam, 2002.

Turvey, Calum G., Xiaolan Xu, Rong Kong, and Ying Cao. "Attitudinal Asymmetries and the Lender-Borrower Relationship: Survey Results on Farm Lending in Shandong, China." *Journal of Financial Services Research* (Springer) 46 (2014): 115-135.

Turvey, Calum Greig. "Credit scoring for agricultural loans: a review with applications." *Agricultural finance review (USA)*, 1991.

West, David. "Neural network credit scoring models." *Computers & Operations Research* (Elsevier) 27 (2000): 1131-1152.

Zech, Lyubov, and Glenn Pederson. "Predictors of farm performance and repayment ability as factors for use in risk-rating models." *Agricultural Finance Review* (MCB UP Ltd) 63 (2003): 41-54.

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

Ziari, Houshmand A., David J. Leatham, and Calum G. Turvey. "Application of mathematical programming techniques in credit scoring of agricultural loans." Proceedings: 1994 Regional Committee NC-207, October 1994, Washington, DC, Regional Research Committee NC-1014: Agricultural and Rural Finance Markets in Transition, 1994.

2.10 Appendices

2.10.1 Logistic Regression

Let $\boldsymbol{x} \in R^n$ be a vector of independent variables of dimension n, and a binary class $y \in \{0, 1\}$, the probability $P\ (y = 1)\ /\ \boldsymbol{x}$, in the logistic regression model for binary classes, is calculated as follows:

$$P\ (y = 1)\ /\ \boldsymbol{x} = \frac{1}{1 - e^{-(w_0 + \boldsymbol{w}^T x)}}$$

The coefficients w_0 and $\boldsymbol{w}$, correspond to the intercept and the vector of parameters associated with the variables **x** respectively. These coefficients can be estimated with the following maximum likelihood function:

$$\min_{w_0,\, \boldsymbol{w}} \sum_{i=1}^{n} y_i\ ln(P\ (y = 1)\ /\ \boldsymbol{x}\ ) + (1 - y_i)\, ln(P\ (y = 1)\ /\ \boldsymbol{x})$$

2.10.2 Neural Networks

A neural network is a set of connected artificial neurons. A neural network functions similarly to the brain, with repeated exposure to a pattern that has stronger association over time.

The multi-layer perceptron neural network (MLP) is the most widely used neural network for classification (Baesens et al., 2003). An MLP consists of an input layer, one or more hidden layers and an output layer, each composed of several neurons. Each neuron processes its inputs and generates an output value that is transmitted to the neurons in the next layer.

The schema of a neural network with a hidden layer can be seen in Figure 3. The neurons of the input layer send signals x_i that are numerical values of attributes or variables, the coefficients W_{ij} correspond to the synaptic weights in the dendrites Of the neuron. Each synaptic weight multiplies its input and defines its relative importance. Each neuron j is activated if the total input exceeds a certain threshold, given by an activation function φ.

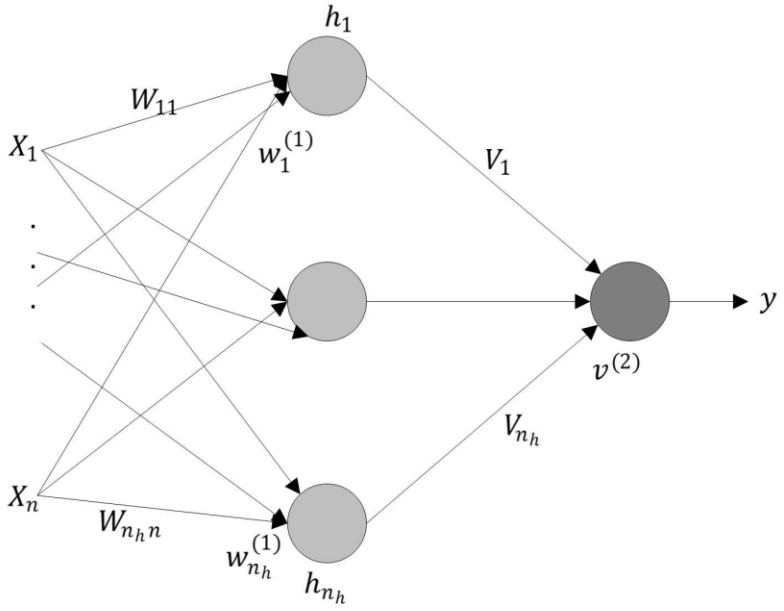

Figure 3: Neural Network

The output h_j of a neuron j and the output y of the final layer are calculated as follows:

$$h_j = \varphi^{(1)}\left(w_j{}^{(1)} + \sum_{i=1}^{n} W_{ji}\, x_i\right)$$

$$y = \varphi^{(2)}\left(v^{(2)} + \sum_{i=1}^{n_h} V_j\, h_j\right)$$

Where $w_j{}^{(1)}$ and $v^{(2)}$ are bias coefficients and$\varphi^{(1)}$ and $\varphi^{(2)}$ are activation functions of the hidden layer and the output layer, respectively; n is the number of input variables x and n_h corresponds to the number of neurons in the hidden layer.

2.10.3 Random Forests

Random Forests correspond to a classifier that combines decision trees in such a way that each uses an independent sample of the data (Breiman, 2001). The training data sets are generated using the bootstrapping method. Bootstrapping corresponds to generating new datasets of the same size using a sampling with replacement. The training of trees with bootstrapping allows to reduce the variance and make the model more robust.

The random Forests algorithm is described below:

- n samples are made with replacement of the data (n is the number of trees to train).

- For each of the samples a classification tree is constructed without pruning. We randomly sample m of the variables and choose the best set of variables.
- The prediction is done by the aggregation of trained trees, where the classification most voted by decision trees is chosen.

2.10.4 Data Statistics

Variable	Average	Standard Dev.	Minimum	Maximum	Percentile 1	Percentile 5	Percentile 10	Percentile 25	Median	Percentile 75	Percentile 90	Percentile 95	Percentile 99
PropertyDistance	33.94	89.69	0.00	2196.55	0.00	0.00	0.00	0.00	16.71	29.02	62.21	108.03	455.57
PropertyLocationN	1.34	1.01	1.00	25.00	1.00	1.00	1.00	1.00	1.00	1.00	2.00	3.00	5.00
Surface	45948.31	226796.56	50.00	80050400.00	400.00	1000.00	1850.00	4360.00	10906.00	31000.00	102000.00	163200.00	600120.00
Cost	668871412.23	2519899961.10	0.00	94077000000.00	40000000.00	56000000.00	83000000.00	154550000.00	322500000.00	670000000.00	1552500000.00	2212500000.00	3570000000.00
Income	28250178725.00	111217971980.00	0.00	22419420000000.00	375000000.00	845000000.00	1425000000.00	3060400000.00	7720000000.00	22132000000.00	61200000000.00	99000000000.00	321330000000.00
CropsNumber	3.07	1.89	1.00	11.00	1.00	1.00	1.00	2.00	3.00	4.00	6.00	7.00	9.00
PropertiesNumber	7.84	9.92	1.00	99.00	1.00	1.00	1.00	2.00	4.00	9.00	20.00	30.00	48.00
IncomeHectare	837632.40	1192376.00	0.00	200000000.00	90000.00	274615.38	365384.62	517289.72	717804.88	968571.43	1233548.39	1500000.00	3213296.40
CostHectare	75972.97	555922.97	0.00	200000000.00	590.91	2545.45	4974.56	12899.63	28333.33	61219.51	128078.82	194545.45	660714.29
IncomeProperty	3574411851.80	7958305030.70	0.00	934142500000.00	177666666.67	350000000.00	504800000.00	921500000.00	1786666666.70	3598500000.00	7602000000.00	11980000000.00	35000000000.00
CostProperty	101318470.72	645498176.40	0.00	26733333333.00	20000000.00	34250000.00	40000000.00	54400000.00	75000000.00	98000000.00	120000000.00	150000000.00	350000000.00
Tenure	0.12	0.33	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00
OfficeClientDist	52.98	157.66	0.00	3490.07	0.00	0.00	0.00	0.00	16.06	29.84	116.64	192.58	817.02
CompanyTime	9.11	5.53	0.00	23.00	0.00	1.00	2.00	5.00	9.00	12.00	17.00	20.00	22.00
ActivityTime	0.98	0.90	0.00	4.00	0.00	0.00	0.00	0.00	1.00	2.00	2.00	3.00	3.00
AccountTime	1.02	0.91	0.00	4.00	0.00	0.00	0.00	0.00	1.00	2.00	2.00	3.00	3.00
PreviousAccountsN	0.00	0.04	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PreviousPurchasesN	42.59	55.24	0.00	472.00	0.00	1.00	3.00	9.00	24.00	54.00	104.00	152.00	260.00
Installment amount	325631.29	549235.75	1.00	34491081.00	5068.00	13079.00	20000.00	58579.00	176115.00	344800.00	732544.00	1169363.00	2477263.00
LoanAmount	443782.47	709330.13	1.00	34491081.00	5973.00	16065.00	26652.00	60765.00	211052.00	586000.00	1040811.00	1577500.00	3516000.00
CreditLimit	77227454.07	162468784.51	0.00	2635000006.00	1.00	3.00	5000000.00	16000001.00	40000000.00	90000000.00	180000007.00	240000002.00	510000005.00
PaymentsNumber	1.34	0.54	1.00	12.00	1.00	1.00	1.00	1.00	1.00	2.00	2.00	2.00	2.00
ProductCategory	36.10	44.40	1.00	99.00	1.00	1.00	1.00	1.00	10.00	99.00	99.00	99.00	99.00
ProductGroupNumber	1.50	0.59	1.00	8.00	1.00	1.00	1.00	1.00	1.00	2.00	2.00	2.00	3.00
ProductNumber	2.20	1.72	1.00	20.00	1.00	1.00	1.00	1.00	2.00	2.00	4.00	5.00	10.00
TotalBalance	6570458.33	16959629.77	0.00	358668412.00	0.00	0.00	39606.00	594855.00	2171795.00	6277972.00	14731374.00	24089684.00	70006214.00
AccountBalance	6569036.87	16959771.76	0.00	358668412.00	0.00	0.00	39512.00	594740.00	2170840.00	6270618.00	14731374.00	24089684.00	70006214.00
RecentAccounts	0.06	0.25	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00
TimeLastAccount	375.29	355.38	0.00	1501.00	0.00	0.00	0.00	0.00	346.00	608.00	889.00	1058.00	1288.00
TimeLastMaturity	122.88	140.17	0.00	1222.00	0.00	0.00	0.00	0.00	55.00	252.00	328.00	353.00	469.00
TimeLastUpdate	352.80	277.58	0.00	1501.00	3.00	23.00	49.00	127.00	287.00	518.00	771.00	909.00	1094.00
current_arrears	150535.16	427661.42	0.00	12237233.00	0.00	0.00	0.00	0.00	0.00	102464.00	452538.00	708770.00	1993245.00
arrears_amount_1_90	148077.07	252587.86	0.00	12143950.00	0.00	0.00	0.00	0.00	65851.00	205647.00	372785.00	555734.00	1155477.00
min_arrears_amount_1_90	14131.28	122021.10	0.00	12143950.00	0.00	0.00	0.00	0.00	0.00	0.00	29767.00	59474.00	220207.00
max_arrears_amount_1_90	668031.12	1102486.08	0.00	14161000.00	0.00	0.00	0.00	0.00	293187.00	801465.00	1774529.00	2773253.00	5516445.00
dec_arrears_amount_1_90	4.50	8.52	0.00	125.00	0.00	0.00	0.00	0.00	1.00	5.00	13.00	20.00	41.00
inc_arrears_amount_1_90	4.84	9.64	0.00	156.00	0.00	0.00	0.00	0.00	0.00	6.00	15.00	23.00	45.00
r_arrears_amount_1_90	0.67	1.60	0.00	64.56	0.00	0.00	0.00	0.00	0.00	0.75	2.04	3.15	6.60
arrears_days_1_90	10.91	17.51	0.00	474.00	0.00	0.00	0.00	0.00	5.29	15.00	27.65	40.00	81.00
min_arrears_days_1_90	3.17	10.98	0.00	474.00	0.00	0.00	0.00	0.00	0.00	0.00	10.00	18.00	48.00
max_arrears_days_1_90	22.94	31.16	0.00	1036.00	0.00	0.00	0.00	0.00	16.00	33.00	55.00	76.00	138.00
timely_payment_1_90	785.96	33601.72	0.00	3599348.00	0.00	0.00	0.00	0.00	0.17	0.60	1.00	1.00	26.49
timely_installment_1_90	0.31	0.36	0.00	1.00	0.00	0.00	0.00	0.00	0.17	0.59	1.00	1.00	1.00
payment_amount_1_90	0.80	0.40	0.00	1.00	0.00	0.00	0.00	0.99	1.00	1.00	1.00	1.00	1.00
n_past_due_1_90	9.96	17.43	0.00	175.00	0.00	0.00	0.00	0.00	3.00	11.00	28.00	44.00	91.00
n_timely_1_90	4.56	8.87	0.00	152.00	0.00	0.00	0.00	0.00	1.00	5.00	13.00	20.00	44.00
arrears_amount_91_180	127570.58	232200.95	0.00	9377517.00	0.00	0.00	0.00	0.00	29812.00	179111.00	351135.00	522828.00	1085472.00
min_arrears_amount_91_180	12394.89	90283.69	0.00	9377517.00	0.00	0.00	0.00	0.00	0.00	0.00	24186.00	56109.00	207743.00
max_arrears_amount_91_180	577534.05	1059134.76	0.00	13915766.00	0.00	0.00	0.00	0.00	117848.00	659734.00	1638812.00	2666115.00	5095060.00
dec_arrears_amount_91_180	3.77	8.04	0.00	127.00	0.00	0.00	0.00	0.00	1.00	4.00	11.00	18.00	39.00
inc_arrears_amount_91_180	4.14	9.29	0.00	156.00	0.00	0.00	0.00	0.00	0.00	4.00	13.00	22.00	44.00
r_arrears_amount_91_180	1.69	61.20	0.00	8995.78	0.00	0.00	0.00	0.00	0.00	0.36	2.09	4.07	16.68
arrears_days_91_180	9.26	16.82	0.00	348.00	0.00	0.00	0.00	0.00	2.21	13.00	25.21	37.89	74.00
min_arrears_days_91_180	2.74	10.46	0.00	348.00	0.00	0.00	0.00	0.00	0.00	0.00	8.00	16.00	46.00
max_arrears_days_91_180	19.40	29.63	0.00	1036.00	0.00	0.00	0.00	0.00	8.00	30.00	52.00	72.00	130.00
timely_payment_91_180	283.62	8656.95	0.00	984843.00	0.00	0.00	0.00	0.00	0.00	0.50	1.00	1.00	1.13
timely_installment_91_180	0.26	0.35	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.50	0.97	1.00	1.00
payment_amount_91_180	0.67	0.47	0.00	1.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00
n_past_due_91_180	8.27	16.18	0.00	166.00	0.00	0.00	0.00	0.00	2.00	9.00	24.00	40.00	82.00
n_timely_91_180	3.74	7.90	0.00	113.00	0.00	0.00	0.00	0.00	0.00	4.00	11.00	18.00	40.00
arrears_amount_181_270	104976.99	218671.39	0.00	9377517.00	0.00	0.00	0.00	0.00	0.00	137599.00	310815.00	462602.00	1027603.00
min_arrears_amount_181_270	10224.39	88908.07	0.00	9377517.00	0.00	0.00	0.00	0.00	0.00	0.00	14044.00	43791.00	171147.00
max_arrears_amount_181_270	469869.27	971677.80	0.00	14161000.00	0.00	0.00	0.00	0.00	0.00	561293.00	1402743.00	2303840.00	4727056.00
dec_arrears_amount_181_270	3.10	7.27	0.00	125.00	0.00	0.00	0.00	0.00	0.00	3.00	9.00	16.00	36.00
inc_arrears_amount_181_270	3.40	8.48	0.00	151.00	0.00	0.00	0.00	0.00	0.00	3.00	11.00	19.00	41.00
r_arrears_amount_181_270	2.71	104.98	0.00	12689.47	0.00	0.00	0.00	0.00	0.00	0.18	1.69	3.49	14.80
arrears_days_181_270	7.64	15.68	0.00	454.00	0.00	0.00	0.00	0.00	0.00	10.00	22.14	34.07	71.00
min_arrears_days_181_270	2.30	9.68	0.00	454.00	0.00	0.00	0.00	0.00	0.00	0.00	6.00	14.00	43.00
max_arrears_days_181_270	15.87	27.70	0.00	454.00	0.00	0.00	0.00	0.00	0.00	24.00	46.00	63.00	123.00
timely_payment_181_270	524.19	31241.10	0.00	3599348.00	0.00	0.00	0.00	0.00	0.00	0.40	0.90	1.00	1.00
timely_installment_181_270	0.22	0.34	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.38	0.86	1.00	1.00
payment_amount_181_270	0.55	0.50	0.00	1.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00
n_past_due_181_270	6.84	15.16	0.00	166.00	0.00	0.00	0.00	0.00	0.00	6.00	21.00	35.00	76.00
n_timely_181_270	3.05	6.89	0.00	99.00	0.00	0.00	0.00	0.00	0.00	3.00	10.00	16.00	34.00