





APPLICATION NOTE



A logistic regression model for consumer default risk

Eliana Costa e Silva ^a, Isabel Cristina Lopes ^b, Aldina Correia ^a and
Susana Faria ^c

^aCIICESI, ESTG, Politécnico do Porto, Felgueiras, Portugal; ^bLEMA, CEOS.PP/ISCAP/P. Porto, Porto, Portugal;

^cDepartment of Mathematics, Centre of Molecular and Environmental Biology, University of Minho, Guimarães, Portugal

ABSTRACT

In this study, a logistic regression model is applied to credit scoring data from a given Portuguese financial institution to evaluate the default risk of consumer loans. It was found that the risk of default increases with the loan spread, loan term and age of the customer, but decreases if the customer owns more credit cards. Clients receiving the salary in the same banking institution of the loan have less chances of default than clients receiving their salary in another institution. We also found that clients in the lowest income tax echelon have more propensity to default. The model predicted default correctly in 89.79% of the cases.

ARTICLE HISTORY

Received 5 February 2019
Accepted 14 April 2020

KEYWORDS

Generalized linear models
logistic regression; default
risk; credit scoring;
applications to actuarial
sciences and financial
mathematics

2010 MATHEMATICS SUBJECT

CLASSIFICATIONS

62-J-12; 91-G-40; 62-P-05

1. Introduction

The objective of this paper is to develop a credit risk prediction model from a small random sample of customers from a Portuguese banking institution.

Credit scoring is the assessment of the risk associated with lending to an organization or an individual [6]. Credit risk modeling, namely its component *Probability of Default* (PD), is very helpful in the consumer credit loan grant decision. A bad customer (*Defaulted*) is commonly taken to be someone who has missed three consecutive months of payments [15]. In fact, three months (or 90 days) of arrears is a standard definition of default at the international level, although it is not the only one. Some countries use 90, 60 or 30 days in arrears as a nonperforming loan definition, and others simply use doubtful or loss loans [3]. Models of credit scoring are based on historical information from a dataset of existing clients, in order to assess whether the prospective client will have a greater chance of being a good or bad payer. Consumer credit risk assessment involves the use of risk assessment tools to manage a borrower's account, from the moment of screening a potential loan application, to the management of the account during its life and possible write-off [6].

Credit scoring is used in almost all forms of consumer lending: credit cards, personal loans, car finance, insurance policies, utility payments. Virtually all major banks use

credit scoring with specialized consultancies providing credit scoring services and offering powerful software to score applicants, monitor their performance and manage their accounts [6]. Financial institution systems incorporate models of credit scoring to permit on-line credit evaluation, and thereby getting higher profits [7].

The Basel Committee on Banking Supervision revised in 2004 the standards governing the capital adequacy of internationally active banks. To evaluate the effects of the Basel II Framework on capital levels, an impact study in 31 countries showed that the minimum required capital levels under the Basel II Framework would on average decrease [2].

The ability of a performance measure to capture the true skill of a model is highly dependent on the data available for assessment [4]. Beyond the social-economical characteristics of the individual, the underlying economic conditions also have a major impact on default. These scoring systems raise social issues, for which institutions are accused of discriminating consumers in the access to credit, and although it is illegal to use some characteristics such as race, sex, or religion, some authors defend the use of surrogate variables.

The existence of correlations in the data used to assess the PD invalidates using statistical tests that require an assumption of independent observations. The logistic regression model provides an appropriate statistical treatment of these correlations [4].

Similar studies have been conducted using logistic regression to assess the credit risk of retail customers (e.g. [9,10,12,16]). Other studies include statistical techniques such as discriminant analysis, linear regression, classification trees, and Bayesian statistics. Also, Operational Research-based approaches, including variants of linear programming, genetic algorithms, nearest neighbor search, and Artificial Intelligence modeling approaches such as neural networks and expert systems have been applied to credit risk prediction [6].

The advantages of using regression models are that it allows to perform statistical tests to identify how important are each of the application form questions to the accuracy of classification, and whether two different questions are essentially asking the same thing and getting equivalent responses. This allows to drop unimportant questions, making scorecards more robust, and helps in deciding what questions to ask in new scorecards [15].

The usual methodology is the lender collecting data from a sample of borrowers who applied, were made an offer of a loan, who accepted the offer and whose subsequent repayment performance has been observed. Information is available on many socio-demographic characteristics (such as income and years at address) of each borrower at the time of application from his/her application form. Typically, information is also collected regarding the repayment performance of each borrower on other loans and of individuals who live in the same neighborhood. A model is parameterized on a training sample, and tested on a holdout sample, to avoid over-parameterization whereby the estimated model fits the nuances in the training sample which are not repeated in the population [6].

In this study, a logistic regression model is applied to credit scoring data from a given financial institution to evaluate the default risk of consumer loans.

The rest of this document is organized as follows. In Section 2, we start by making a brief introduction to logistic regression. In Section 3, the data structure used in this work is detailed, followed by the exploratory analysis of all the variables. Next, in Section 4, we build the logistic regression model for default risk, test for interactions between variables, and present estimates of the selected model. The model validation is presented in Section 5,

where goodness-of-fit tests and residuals analysis are presented. Finally, in Section 6, some conclusions are drawn and an outlook for future work is presented.

2. Logistic regression

When the response variable Y follows a Bernoulli distribution of parameter μ , then the generalized linear model uses the *logit function* as the canonical link function and becomes a *logistic regression model*. As $Y_i \sim \text{Ber}(\mu_i)$, then $\mu_i = P(Y_i = 1)$.

The variable `Default` is a binary variable Y such that $Y = 1$ if defaulted, and 0 otherwise. Using the logistic regression model, the PD is a function of a set of explanatory variables \mathbf{X} as follows:

$$P(Y = 1|\mathbf{X}) = \frac{1}{1 + e^{-\beta\mathbf{X}}}. \quad (1)$$

To estimate the regression coefficients of the GLM models, the maximum likelihood method is used. The implementation provided by the command `glm` from R is used. The estimates for β are obtained as solution of a system of likelihood equations, that is usually solved using the Nelder and Wedderburn algorithm, which is an iterative method that uses Fisher's information matrix. Note that several methods may be used to estimate the coefficients of a GLM model (e.g. Bayesian methods and M-estimation).

3. Data description

The dataset contains financial data regarding consumer loans and a brief social characterization of the clients of a Portuguese banking institution, between January 2008 and December 2009, where the official currency is Euro. It is composed of 14 variables, of which eight are quantitative and six are qualitative:

- Quantitative variables:
 - `Contracted Capital`: represents the capital negotiated in the loan agreement (in Euros).
 - `Capital Outstanding`: represents the capital that it is still owed to the bank at the moment (in Euros).
 - `Spread`: The interest rate spread is the interest rate charged by banks on loans to private sector customers minus the interest rate paid by commercial banks for demand, time, or savings deposits (in per cent points). This is an individual spread of each loan, which is established in the loan contract.
 - `Term`: the length of the loan (in years).
 - `Monthly Installment`: the amount to be paid to the bank each month (in Euros).
 - `Age`: age of the borrower (in years).
 - `Seniority`: number of years that the borrower has been a client of the banking institution (in years).
 - `Credit Cards`: number of credit cards owned by the client.
- Qualitative variables:
 - `Sex`: gender of the borrower. It is a binary variable coded as F for female and M for male.

- **Marital Status:** marital status of the client, coded in the following way: 0 = Unknown; 1 = Single; 2 = Union of Fact; 3 = Married without community property; 4 = Married in community of purchased property; 5 = Separated; 6 = Divorced; 7 = Widowed.
- **Salary:** indicates whether the salary of the borrower is received in an account in the same banking institution in which the loan is made or if it is received in some other banking institution. It is a binary variable coded as: 1 = The salary is received in an account in this same banking institution; 0 = The salary is received in an account in other banking institution.
- **Other Credit:** indicates if the client has other credits. It is a binary variable coded as 1 = Yes, 0 = No.
- **Tax Echelon:** the IRS tax echelon of the client, where 1 is the lowest income echelon, and 6 is the highest income echelon.
- **Default:** indicates if the borrower is in default, i.e. if the borrower has not made a scheduled payment of interest or principal. It is a binary variable coded as 0 = No, 1 = Yes.

This dataset is a simple random sample of all the banking institution records, composed of 3221 individuals, where 319 defaulted, making an observed default rate of 10%.

The dataset has eight quantitative explanatory variables (**Contracted Capital**; **Capital Outstanding**; **Spread**; **Term**; **Monthly Installment**; **Age**; **Seniority**; **Credit Cards**). The first seven are continuous and the last is discrete. For each variable, two groups will be considered according to the variable **Default** (one group when **Default** is 0 and another when **Default** is 1).

In addition, the dataset has five qualitative variables: three of them are binary (**Sex**, **Salary** and **Other Credit**), **Marital Status** is a qualitative nominal variable, and **Tax Echelon** is a qualitative ordinal variable.

In the years 2008 and 2009, Portugal was in a favorable macroeconomic environment. In this period, the end of an economic growth cycle was observed, with the Gross Domestic Product per capita having reached 16,942 Euros in 2008 (Source: INE¹ – Gross domestic product per capita at current prices – Base 2011). The inflation rate was in sharp decline, from 2.6% in 2008 to a negative inflation rate in 2009 of –0.8% (Source: INE – Consumer price index – average rate of change over the last 12 months – Base 2012), reflecting a time of economic expansion in the country. In 2008, the unemployment rate stood around 8.4% and 9.5%, having experienced a slight reduction in 2008 compared to previous years, but in 2009 it started to increase, achieving 11.5% in the end of the year (Source: INE – Unemployment rate (%) of the active population aged between 15 and 74 years old). In the following years, there was a big increase in the unemployment rate due to the crisis that hit Portugal in the years 2011–2012.

Previously, in [5], the authors analyze, using statistical inference techniques, such as the Mann–Whitney–Wilcoxon and Pearson Chi-squared independence tests, to infer the factors that influence credit risk from a small random sample of customers from a Portuguese banking institution. The nonparametric Mann–Whitney–Wilcoxon test was used to compare the medians of each variable and the results show that, for a 5% significance level, there are differences between the medians in the groups of defaulters and non-defaulters, for the variables **Contracted Capital**, **Capital Outstanding**, **Spread**, **Age** and

Credit Cards. Also, when testing the variables Term, Monthly Installment and Seniority there is no statistical evidence, at a 5% significance level, to reject the null hypothesis. Hence these variables may not be relevant to explain the variable Default.

Pearson Chi-squared independence test was used to check if the qualitative variables have some influence on the probability of occurring a default. The results show that the credit default risk depends on receiving the salary in the same banking institution. The results also show that, at a 5% significance level, the tax echelon is not independent of the default risk. In fact, for clients in the lowest tax echelon, the percentage of defaulted clients is larger, comparing to what happens in other echelons.

In summary, the authors found that the decision of granting consumer credit should take into account variables such as Capital Outstanding, Spread, Age, Credit Cards, Salary and Tax Echelon. While demographic variables such as Sex and Marital Status, economic characteristics of the individuals such as Seniority and Other Credit, and characteristics of the account such as Term and Monthly Installment, were found not to be relevant to act on default.

4. Logistic regression model

For building the logistic regression model, a simple random sample of 80% of the records was considered. First, a logistic regression model was fit to the sample of 2577 records, and then this model was applied to the entire original dataset, consisted of 3221 records, to predict the variable Default. The variable Tax Echelon, with only 5 categories, was used as a cofactor, instead of the original variable Tax Echelon, as explained in Section 3.

4.1. Building a logistic regression model

Several logistic regression models for predicting the default risk were tested. For the selection of the most suitable model, the likelihood ratio test and the AIC (Akaike Information Criterion) were used.

The variables for which the null hypothesis of the Wald test is rejected, at a significance level of 5%, and therefore are significant covariables in the model, are: Spread, Term, Age, Credit Cards, Salary and Tax Echelon. The summary of this model is presented in Table 1 and in Equation (2).

$$\log \left(\frac{\mu}{1 - \mu} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9, \quad (2)$$

where μ is the mean of variable Default (and represents the PD), X_1 is the variable Spread, X_2 is the variable Term, X_3 is the variable Age, X_4 is the variable Credit Cards, X_5 is a dummy variable for when Salary = 1, X_6 is a dummy variable for when Tax Echelon = 2, X_7 is a dummy variable for when Tax Echelon = 3, X_8 is a dummy variable for when Tax Echelon = 4, X_9 is a dummy variable for when Tax Echelon = 5.

The resulting model confirms most of the conclusions obtained in Section 3. All the variables suggested by the exploratory analysis were found to be significant in the model.

Table 1. Summary of the logistic regression model obtained with Equation (2).

Coefficients	Estimate	Std. Error	z	value	p-value
(Intercept)	−4.293	0.951	−4.513		6.40e−06***
Spread	0.352	0.103	3.427		0.001***
Term	0.042	0.013	3.121		0.002**
Age	0.043	0.018	3.360		0.001***
Credit Cards	−1.550	0.225	−6.884		5.84e−12***
factor(Salary)1	−0.842	0.154	−5.473		4.42e−08***
factor(Tax Echelon)2	−3.235	1.010	−3.203		0.001**
factor(Tax Echelon)3	−3.367	0.716	−4.700		2.61e−06***
factor(Tax Echelon)4	−2.556	0.514	−4.975		6.54e−07***
factor(Tax Echelon)5	−4.636	1.004	−4.619		3.86e−06***
Null deviance:	1654.7 on 2576 degrees of freedom				
Residual deviance:	1183.2 on 2567 degrees of freedom				
AIC:	1203.2				

Note: ** <1% and *** <0.1%.

Only the variable `Term`, that was not suggested by the exploratory analysis to be relevant, is now found to be relevant too.

4.2. Testing for interactions between variables

We considered interactions between the quantitative and qualitative variables present in the best model found in the previous section. After many experiences, the only interaction that always came out significant, according to the Wald test, for a significance level of 5%, was the interaction between `Age` and `Credit Cards`. We also considered several models with interactions between the variables in Equation (2) and the variables that were previously removed, but no other interaction came out relevant.

This last model has a smaller AIC than the AIC of the model in Equation (2), which indicates that this last model would be preferable to the previous one.

The equation of this model, considering the interactions between `Age` and `Credit Cards` is the following:

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_3 * X_4, \quad (3)$$

where μ is the mean of variable `Default`, X_1 is the variable `Spread`, X_2 is the variable `Term`, X_3 is the variable `Age`, X_4 is the variable `Credit Cards`, X_5 is a dummy variable for `Salary = 1`, X_6 is a dummy variable for `Tax Echelon = 2`, X_7 is a dummy variable for `Tax Echelon = 3`, X_8 is a dummy variable for `Tax Echelon = 4`, X_9 is a dummy variable for `Tax Echelon = 5`.

We also performed the Likelihood Ratio Test between these two nested models. The difference between the deviances of both models is 14.374, with 1 degree of freedom, producing a p -value of approximately 0, which causes the null hypothesis of this test to be rejected at a 1% level of significance.

In conclusion, the model with the interaction between `Age` and `Credit Cards` is preferable to the model without this interaction.

Table 2. Estimates for the coefficients of the logistic regression model in Equation (3).

	β	$\exp(\beta)$	95% CI for $\exp(\beta)$	
			LCI	UCI
(Intercept)	-4.037	0.018	0.003	0.111
Spread	0.347	1.415	1.157	1.726
Term	0.043	1.044	1.018	1.073
Age	0.036	1.037	1.011	1.064
Credit Cards	-5.607	0.004	0.000	0.035
factor(Salary)1	-0.825	0.438	0.327	0.591
factor(Tax Echelon)2	-3.231	0.040	0.002	0.180
factor(Tax Echelon)3	-3.366	0.035	0.006	0.110
factor(Tax Echelon)4	-2.541	0.080	0.024	0.190
factor(Tax Echelon)5	-4.621	0.010	0.001	0.044
Age:Credit Cards	0.084	1.088	1.041	1.141
Null deviance:	1654.7 on 2576 degrees of freedom			
Residual deviance:	1168.9 on 2566 degrees of freedom			
AIC:	1190.9			

In the following sections, the model presented in Equation (3) will be validated and used to predict the credit default risk.

4.3. Model estimates

In Table 2, the estimates for the model parameters are shown.

The standard errors of the parameters are the square root of the main diagonal of the inverse of Fisher's information matrix, which contains the covariances of the parameters.

In logistic regression models, rather than looking at the coefficients β_i per se, it is more important to focus on the values of $\exp(\beta_i)$, because they represent the influence that the increase in an independent variable X_i has in the probability of the dependent variable Y becoming 1.

It follows that:

$$\log \left(\frac{P(Y = 1|X_i)}{1 - P(Y = 1|X_i)} \right) = \beta_0 + \beta_1 X_1 + \dots + \beta_i X_i + \dots + \beta_p X_p \quad (4)$$

$$\Leftrightarrow \frac{P(Y = 1|X_i)}{1 - P(Y = 1|X_i)} = e^{\beta_0 + \beta_1 X_1 + \dots + \beta_i X_i + \dots + \beta_p X_p} \quad (5)$$

The term on the left side of the Equation (5) is called the *odds* of the variable Y . In our model, it represents the ratio between the probability of a client committing default and the probability of not committing default.

The *Odds Ratio* (OR) is a ratio of two odds. The OR between the odds of Y_2 , which is Y given that the set of covariates is $X = X_2$, and the odds of Y_1 , which is Y given that the set of covariates is $X = X_1$, is the following:

$$OR = \frac{Odds_2}{Odds_1} = \frac{\frac{\mu_2}{1-\mu_2}}{\frac{\mu_1}{1-\mu_1}} = \frac{\frac{P(Y_2=1)}{P(Y_2=0)}}{\frac{P(Y_1=1)}{P(Y_1=0)}}. \quad (6)$$

If X and Y are independent, $OR = 1$ is the baseline for comparison [1]. If in Equation (6) we obtain a value $OR > 1$, then the odds of default are higher when $X = X_2$ than when $X = X_1$.

If one of the quantitative independent variables, X_i , is increased in one unit, while the remaining variables are maintained constant, then the odds are given by

$$\frac{P(Y = 1|X_i + 1)}{1 - P(Y = 1|X_i + 1)} = e^{\beta_0 + \beta_1 X_1 + \dots + \beta_i (X_i + 1) + \dots + \beta_p X_p} \quad (7)$$

$$\Leftrightarrow \frac{P(Y = 1|X_i + 1)}{1 - P(Y = 1|X_i + 1)} = e^{\beta_i} \times e^{\beta_0 + \beta_1 X_1 + \dots + \beta_i X_i + \dots + \beta_p X_p} \quad (8)$$

$$\Leftrightarrow Odds(Y|X_i + 1) = e^{\beta_i} \times Odds(Y|X_i). \quad (9)$$

This means that $\exp(\beta_i) = (Odds(Y|X_i + 1))/Odds(Y|X_i)$ represents the OR. The estimates for the coefficients β_i of the logistic regression model in Equation (3) are presented in Table 2, along with the computations of $\exp(\beta_i)$ and the 95% confidence intervals for $\exp(\beta_i)$.

The estimates for the coefficients of the variables *Spread*, *Term*, *Age* and *Age*Credit Cards* are positive, which causes that $\exp(\beta)$ in these cases is greater than 1, meaning that an increase in one of these variables would reflect in an increasing chance of defaulting. For example, for the variable *Spread*, $\exp(\beta_1) = 1.415$, which states that for each percent point increased in the spread of a loan (and maintaining the rest of the variables constant), the OR of defaulting increases 41.5%. Similarly, if the term of the loan agreement is extended in one year, the odds of the client committing a default increase 4.4%. For two clients with exactly the same spread, term, number of credit cards, salary and tax echelon, but with an age difference of one year, the older client has 3.7% more chances of defaulting than the younger client.

The value of $\exp(\beta)$ of *Credit Cards* is 0.004, and the corresponding 95% confidence interval lies between 0.000 and 0.035, which is completely situated below 1, hence it means the OR are significantly different from one another at 5% level of significance, and that the more credit cards a person has, the less is the PD. In fact, it means that for each extra credit card, the OR of default decreases 99.6%.

However, the variables *Age* and *Credit Cards* combined have $\exp(\beta) = 1.088 > 1$, which means that when the product of these two variables increases in one unit and the remaining variables are left constant, the OR of defaulting increases 8.8%.

The value of $\exp(\beta)$ for the binary variable *Salary* is 0.438, which means that a client that receives her/his salary in the same bank of the loan (*Salary* = 1) has 56.2% less chances of defaulting than a client that receives the salary in another institution (*Salary* = 0).

For the variable *Tax Echelon*, four dummy variables were created, with *Tax Echelon* = 1 as the reference category. All the coefficients of these dummy variables are such that $\exp(\beta) < 1$. This represents that all these tax echelons (2, 3, 4 and 5) have less chances of defaulting than the reference (*Tax Echelon* = 1). For example, if two clients have the same loan conditions but one is in *Tax Echelon* = 1 and the other is in *Tax Echelon* = 2, the latter has 96% less chances of defaulting.

5. Model validation

The final logistic regression model was the model in Equation (3), for which the coefficient estimates are in Table 2. Before using this model to estimate the probability of a client of the bank defaulting, the model has to be validated through a series of statistical tests, and the assumptions of the model have to be verified.

5.1. Goodness-of-fit tests

An important topic in modeling exercise is the goodness-of-fit test: testing the null hypothesis that the model fits the data well versus the opposite. The goodness-of-fit of a binary logistic model can be done using the Hosmer–Lemeshow test. This test can easily be obtained using the output from several statistical packages and along with the Pearson's chi-square test are commonly recommended for assessing lack of fit for proposed logistic regression models. The Hosmer–Lemeshow test is performed by sorting the n observations by the predicted probabilities, and forming g groups with approximately the same number of subjects in each group (m). Then, the test statistic is calculated as

$$C_g = \sum_{j=1}^g \frac{(e_j - o_j)^2}{m\bar{e}_j(1 - \bar{e}_j)}, \quad (10)$$

where e_j is the sum of the estimated success probabilities of the j th group while o_j is the sum of the observed success items of the j th group, and the term \bar{e}_j is the mean of the estimated success probabilities of the j th group. It is known that under the null hypothesis, C_g obeys a chi-square distribution $\chi_{(g-2)}^2$. In practice, the number of groups g is usually chosen to be 10. In the final model, the Hosmer–Lemeshow test reported a p -value of 0.765 and did not indicate lack of fit.

5.2. Residuals analysis

The model may also be validated by studying the residuals and performing regression diagnostics. Regression diagnostics are certain quantities computed from the data with the purpose of pinpointing influential points and study their impact on the model and the subsequent analysis [13]. Once identified, these influential points can be removed or corrected.

For a logistic regression model, Pearson residuals are defined as

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{v}_i}}, \quad (11)$$

where $\hat{v}_i = \hat{\mu}_i(1 - \hat{\mu}_i)$, and deviance residuals are computed as

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i) \sqrt{2 \left(y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) + (1 - y_i) \log \left(\frac{1 - y_i}{1 - \hat{\mu}_i} \right) \right)}. \quad (12)$$

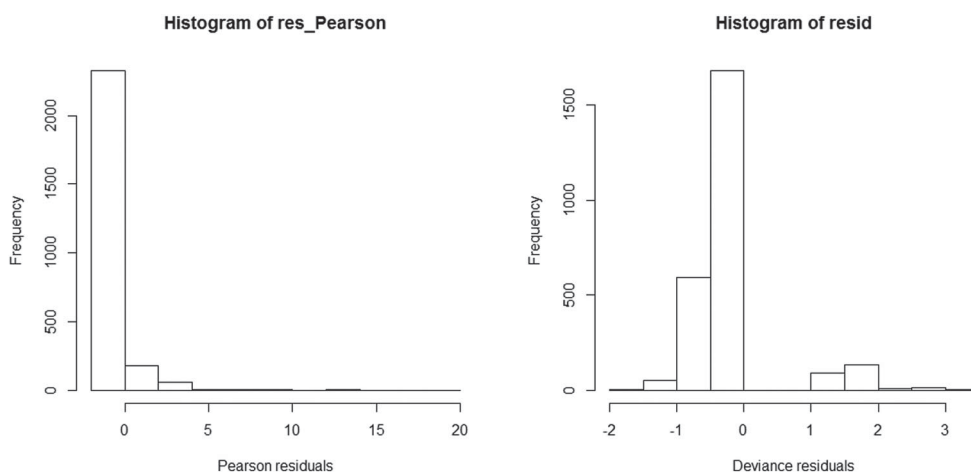


Figure 1. Histograms of the Pearson residuals (mean: 0.004; variance: 0.952) and Deviance residuals (mean: -0.106 ; variance: 0.445) for the 2577 individuals.

The standardized Pearson residual r_i^{PS} and standardized deviance residuals r_i^{DS} are defined as

$$r_i^{PS} = \frac{r_i^P}{\sqrt{1 - h_{ii}}}, \quad r_i^{DS} = \frac{r_i^D}{\sqrt{1 - h_{ii}}}, \quad (13)$$

where h_{ii} is the i th leverage value, which is, in fact, the i th diagonal element of the leverage matrix

$$H = V^{1/2}X(X^TVX)^{-1}X^TV^{1/2} \quad (14)$$

and V is a diagonal matrix with diagonal element v_i .

Figure 1 shows that, as expected, the residuals do not have a standard normal distribution. In fact, the distribution, for both residuals, is asymmetric.

On the other hand, for the deviance residuals, Figure 2 reveals several outliers. However, only 26 observations (approximately 1% of the total of observations) have deviance residuals larger than 2 in absolute value, i.e. $|r_i^D| > 2$. Therefore most of the residuals are between -2 and 2 . The conclusion is also that the model is adequate.

Outliers are observations corresponding to exceptionally large residuals and are examples of atypical observations. High leverage points and influential observations are also atypical observations. High leverage points are points in remote regions with high influence in the adjusted value. Observations whose presence (or absence) can make a huge impact on the fitting of the model are called influential.

Note that influential observations need not be outliers. In fact, as stated in [8], outliers are not always highly influential. Also leverage points need not be influential, and influential observations are not necessarily high leverage points.

The leverage value h_i of observation y_i represents the weight of the observation y_i in the adjusted value. Typically, the observation y_i is considered a high leverage point if $h_i > 2(p/n)$ or if $h_i > 3(p/n)$, where p is the number of parameters in the model and n is the number of observations. Figure 3 shows the leverage values for all the 2577 observations. On the left, it is also plotted the line $2(p/n)$, and on the right, it is plotted the line $3(p/n)$.

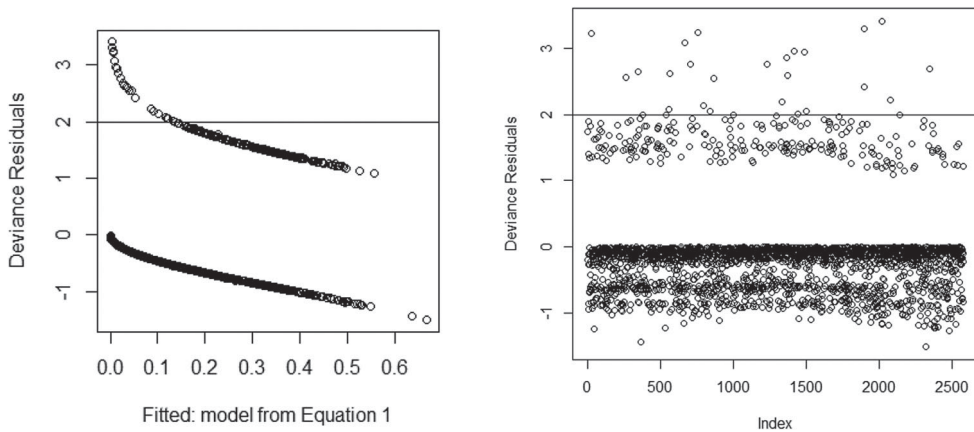


Figure 2. Fitted values and index versus residual deviance.

Assuming that a point is a high leverage one if $h_i > 2(p/n)$, there are 258 (10%) high leverage points, with maximum, mean and minimum 0.067, 0.014 and 0.009, respectively. If the criterion is $h_i > 3(p/n)$, then 87 (3.38%) observations are high leverage, with maximum, mean and minimum 0.067, 0.022 and 0.013, respectively.

In [8], Imon and Hadi enhance the importance of the identification of high leverage points, since they greatly affect the fitted values and consequently might cause problems such as erroneous goodness-of-fit statistics, wrong OR, wrong Wald statistics, etc.

A measure of the influence of each observation on the regression parameter estimates is Cook's distance. For a logistic regression model, the i th Cook's distance, c_i , is defined as

$$c_i = (p + 1)^{-1} (\hat{\beta} - \hat{\beta}^{(-i)})^T (X^T V X) (\hat{\beta} - \hat{\beta}^{(-i)}), \quad (15)$$

where $\hat{\beta}^{(-i)}$ is the estimated parameter of β with the i th observation deleted. To avoid fitting the model $(n + 1)$ times, we shall use the usual approximation to Cook's distance

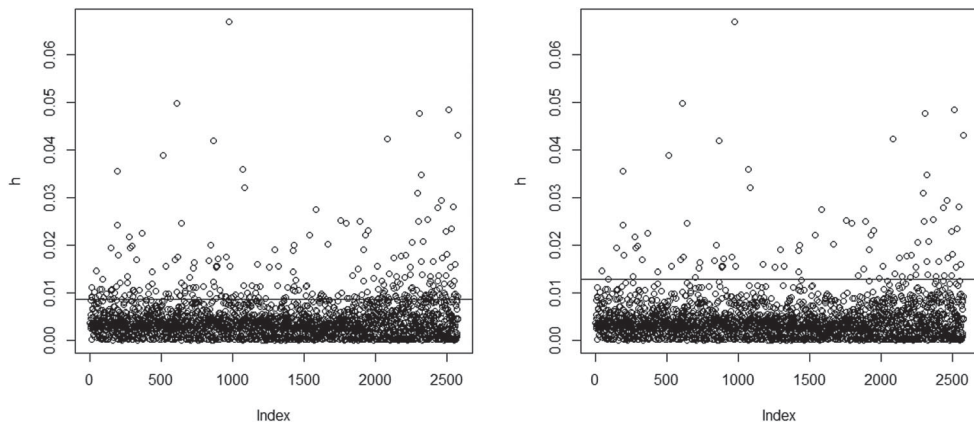


Figure 3. Leverage level considering that a point is a high leverage one if $h_i > 2(p/n)$, on the left, or $h_i > 3(p/n)$, on the right.

Table 3. Confusion matrix, considering only the sample used to create the logistic regression model.

Default	Observed		Total
	0	1	
predicted			
0	2313	251	2564
1	11	2	13
Total	2324	253	2577

Table 4. Confusion matrix of the logistic regression model when applied to all the available data.

Default	Observed		Total
	0	1	
predicted			
0	2889	316	3205
1	13	3	16
Total	2902	319	3221

given by

$$c_i = \frac{(r_i^P)^2}{p + 1} \times \frac{h_{ii}}{1 - h_{ii}}; \tag{16}$$

it combines leverage and residuals. It is common practice to plot c_i against i (see Figure 4). An observation y_i is considered to be influential if $c_i > F_{p,n-p}(0.5)$. In our model, none of the 2577 observations is an influential one.

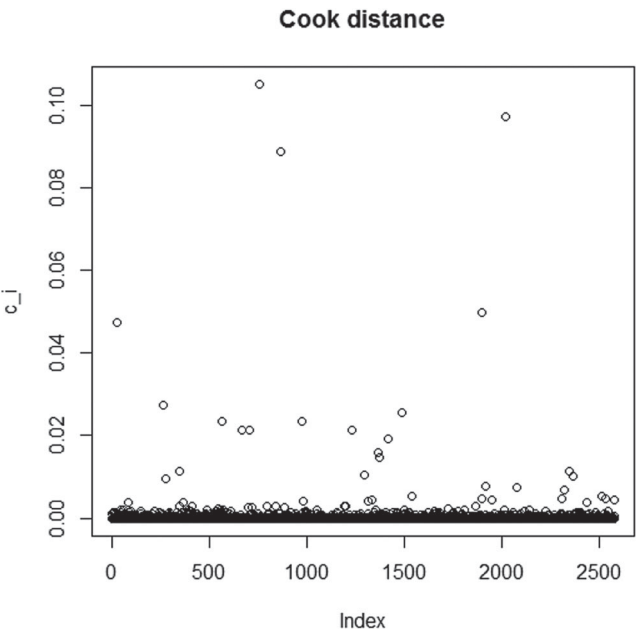


Figure 4. Cook's distance.

6. Conclusions

We developed a logistic regression model to predict the default credit risk using data from consumer loans in a banking institution in Portugal.

The explanatory variables found relevant were Spread, Term, Age, Credit Cards, Salary and Tax Echelon. It was found that the risk of default increases with the loan spread, loan term and age of the customer, but decreases if the customer owns more credit cards. Clients receiving the salary in the same banking institution of the loan have less chances of default than clients receiving their salary in another institution. We also found that clients in the lowest income tax echelon have more propensity to default.

The fitted values for Default, obtained when the logistic regression model is applied to the whole dataset, are depicted in Figure 5.

The model was validated in terms of goodness-of-fit, residuals analysis, and lack of influential points.

The classification table in Table 3 is called a *confusion matrix*. The PD was estimated using Equation (1). The predicted value of Default was rounded to 1 if the PD estimated from the model was ≥ 0.5 , and rounded to 0 otherwise. The percent of correctly classified cases is $(2313+2)/2577 = 89.83\%$.

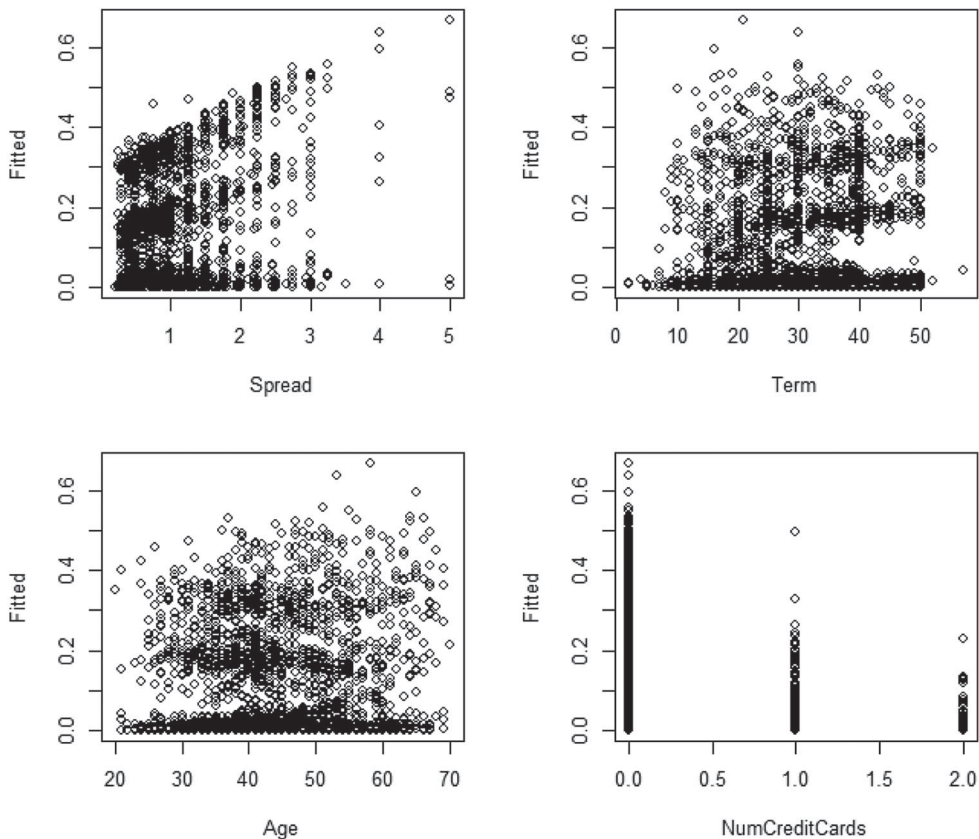


Figure 5. Fitted values for Default (when model is applied to the whole dataset).

The model was also applied to a larger dataset, for which the confusion matrix is in Table 4.

There were 13 cases where the model estimated the default to happen, but in reality it did not, and 316 cases where default occurred in fact, but the model estimated that it would not. The rate of correctly classified cases is $(2889 + 3)/3221 = 89.79\%$. Our rate is better than the rates obtained with logistic regression that appeared in most papers (as cited in [15] and in [6]), except for the rates presented in [11].

As said before, the cutting point for considering $\text{Default} = 1$ was 0.5. For understanding what would happen if different cutting points were considered, the *receiver operating curve* (ROC), presented in Figure 6, is commonly used. The ROC curve is a plot of the proportion of bads classified as bad (*Sensitivity*), against the proportion of goods classified as bad ($1 - \text{Specificity}$), at all values of the cutting point [6].

For validation of the model, we used the traditional Chi-square and Hosmer–Lemeshow tests.

Our link function was validated, but a different link function could also be used in the generalized linear model, as for example the *probit* model, which relies on the probit link function, transforming the probability of default μ to z-scores of the standard normal distribution [1].

Although the logistic regression model is widely used, survival analysis models have recently been found better to build consumer credit models [14], because they provide not only the expected default time, but also the PD of each time point in the future. This is important because a loan default near the end of the loan term may bring a profit greater than its cost before it defaults [17]. Survival models could not be used in this study, because the precise time of the default was not available in our dataset.

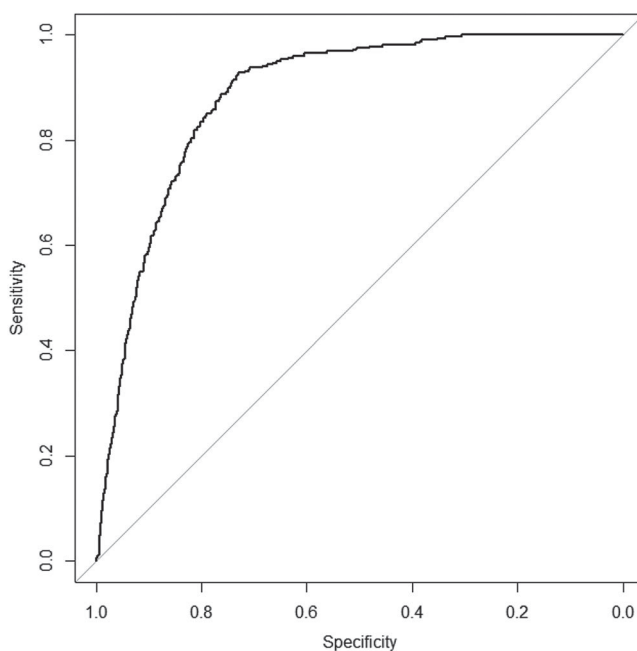


Figure 6. ROC curve (when model is applied to the whole dataset).

Note

1. Instituto Nacional de Estatística (INE) – Statistics Portugal, available at www.ine.pt.

Acknowledgments

This work has received funding from FEDER funds through P2020 program and from national funds through FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) under the project UID/GES/04728/2020.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work has received funding from FEDER funds through P2020 program and from national funds through FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) under the project UID/GES/04728/2020.

ORCID

Eliana Costa e Silva  <http://orcid.org/0000-0001-9757-6687>

Isabel Cristina Lopes  <http://orcid.org/0000-0002-4833-470X>

Aldina Correia  <http://orcid.org/0000-0002-4693-4867>

Susana Faria  <http://orcid.org/0000-0001-8014-9902>

References

- [1] A. Agresti, *An Introduction to Categorical Data Analysis*, 2nd ed., Wiley, Hoboken, NJ, 2007.
- [2] Basel Committee on Banking Supervision, *Results of the fifth quantitative impact study (QIS 5)*, Bank for International Settlements, Basel, Switzerland, 2006.
- [3] R. Beck, P. Jakubik, and A. Piloju, *Key determinants of non-performing loans: New evidence from a global sample*, *Open Econ. Rev.* 26 (2015), pp. 525–550.
- [4] J. Chen, S. Dhar, D. Duffy, Y. Liu, R.O. Moore, M. Pedneault, A. Pole, Y.Q. Qian, D. Rumschitski, T. Wang, and M. Zyskin, *New performance measures for credit risk models*, Tech. Rep., Technical report delivered to Standard & Poor's Rating Services following 2014 Workshop on Mathematical Problems in Industry, held at New Jersey Institute of Technology, Newark, NJ, USA, 2014.
- [5] E. Costa e Silva, I.C. Lopes, A. Correia, and S. Faria, *Consumer default risk assessment in a banking institution*, *AIP Conf. Proc.* 1790 (2016), p. 140007.
- [6] J.N. Crook, D.B. Edelman, and L.C. Thomas, *Recent developments in consumer credit risk assessment*, *Eur. J. Oper. Res.* 183 (2007), pp. 1447–1465.
- [7] M.A. Gouvêa and E.B. Gonçalves, *Credit risk analysis applying logistic regression, neural networks and genetic algorithms models*, Production and Operations Management Society 18th Annual Conference, Dallas, Texas, USA, 2007.
- [8] A.H.M.R. Imon and A.S. Hadi, *Identification of multiple high leverage points in logistic regression*, *J. Appl. Stat.* 40 (2013), pp. 2601–2616.
- [9] M. Karan, A. Ulucan, and M. Kaya, *Credit risk estimation using payment history data: A comparative study of Turkish retail stores*, *Cent. Eur. J. Oper. Res.* 21 (2013), pp. 479–494.
- [10] H.H. Kuangnan Fang, *Variable selection for credit risk model using data mining technique*, *J. Comput.* 6 (2011), pp. 1868–1874.

- [11] E.K. Laitinen, *Predicting a corporate credit analyst's risk estimate by logistic and linear models*, Int. Rev. Financ. Anal. 8 (1999), pp. 97–121.
- [12] S. Mestiri and M. Hamdi, *Credit risk prediction: A comparative study between logistic regression and logistic regression with random effects*, Int. J. Manage. Sci. Eng. Manage. 7 (2012), pp. 200–204.
- [13] A. Nurunnabi, A.R. Imon, and M. Nasser, *Identification of multiple influential observations in logistic regression*, J. Appl. Stat. 37 (2010), pp. 1605–1624.
- [14] S. Privara, M. Kolman, and J. Witzany, *Recovery rates in consumer lending: Empirical evidence and model comparison*, Tech. Rep., SSRN, 2013. Available at <http://ssrn.com/abstract=2343069>.
- [15] L.C. Thomas, *A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers*, Int. J. Forecast. 16 (2000), pp. 149–172.
- [16] S. Westgaard and N. van der Wijst, *Default probabilities in a corporate bank portfolio: A logistic model approach*, Eur. J. Oper. Res. 135 (2001), pp. 338–349.
- [17] C. Yeh and T. Lee, *Credit card behavior tells the risk of unsecured consumer credit loan – application of survival table*, Int. Res. J. Appl. Finance IV (2013), pp. 730–739.