Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 170 (2020) 1128–1133

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Statistical Methods and Artificial Intelligence (IWSMAI)
April 6-9, 2020, Warsaw, Poland

Modeling car loan prepayment using supervised machine learning

Sara Zahi[a]*, Boujemâa Achchab[a]

[a]*Laboratory of Systems Modelization and Analysis for Decision Support, National School of Applied Science, Hassan 1st University, Berrechid, Morocco*

Abstract

Logistic regression is a widely used machine-learning model for predicting categorical outcomes. It uses a number of explanatory variables to predict the values of the target variable that can be either binomial or multinomial. It is used in a number of fields such as cancer detection problems, risk modeling and any subject that requires computing the probability of an event occurrence. In this paper, we propose to apply this supervised machine-learning model to study prepayment risk and its determinants concerning car loans based on a number of characteristics.

Keywords: supervised learning; logistic regression; prepayment risk;

1. Introduction

In many countries, one of the ways that individuals finance their purchases is by subscribing to bank loans. Banking institutions provide their clients with various types of loans such as mortgage loans, car loans and consumer loans. Due to the nature of the transactions these financial institutions make, they can face various types of risks. “These risks include interest rate risk, prepayment risk and credit risk” [1]. Whilst prepayment risk refers to “the risk that a borrower will pay off the loan before maturity” [1], it can cause the banking institutions loss of interests and often forces them to proceed to the reinvestment of the prepaid amount at a lower rate of return. Despite the

* Corresponding author. Tel.: + 212-522-324-758 ; fax: + 212-522-534-530
E-mail address: s.zahi@uhp.ac.ma

10.1016/j.procs.2020.03.055

consequences that prepayment can have on banking institutions, credit-scoring studies have mainly focused on modeling default risk and rarely on prepayment risk [2]. Therefore, to help banking institutions face the risk that a loan will be repaid earlier than the originally agreed termination date, it is essential to provide them with aid decision-making tools. In this paper, we will focus on car loan prepayment and build a model that is able to predict whether a subscriber is likely to pay back the loan before maturity or not. We will apply logistic regression which is a supervised machine-learning algorithm used for predicting the probability of an event occurrence. This will help banks minimize the loss in profit caused by prepayment risk.

Machine-learning is the science that allows computers to find patterns within data and construct classification and prediction models whether well labeled data are used (supervised learning) or not (unsupervised learning). Supervised learning is "the most common form of machine learning" [3]. Using it requires to "specify the output" [4]. It is the type of learning where the training uses well-labeled data. Afterwards, a new set of data, called the test set is introduced to the model that produces an outcome. Logistic regression is a type of supervised learning that mainly makes predictions regarding issues where the estimate of a probability as an output is required. Since logistic regression is used to find the probability of any event occurring, we aim to use this popular algorithm to model the prepayment risk of car loans.

This paper is structured as follows: after the introduction, we will present the research methodology in the second section and the supervised machine-learning model used. In the third section, we will present the results and discussion for our case study before concluding.

2. Research methodology

2.1. Supervised learning

In supervised learning, "training data that is provided to the algorithm, contains the desired solutions called labels" [5]. Therefore, supervised learning models are trained on a labeled dataset where we find both the input and the output parameters. The figure (Fig. 1.) below illustrates the difference between supervised and unsupervised learning:

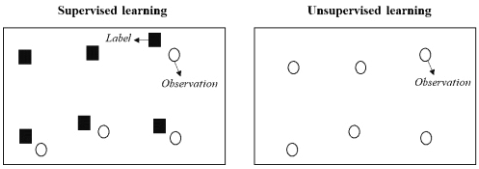

Fig. 1. Supervised learning vs unsupervised learning

2.2. Logistic regression

Logistic regression is a supervised machine-learning algorithm that analyzes a dataset containing one or multiple variables in order to determine an outcome. It is classified among supervised learning models because it is a classification algorithm that is trained on a labeled dataset and uses true labels during its training phase. "It is often used to estimate the probability that an observation belongs to a particular class" [5]. "It predicts the probability of occurrence of an event by fitting data to a logit function" [6]. The difference between binary logistic regression and multinomial logistic regression is that the dependent variable of the binary logistic regression has two outcomes while the dependent variable of the multinomial logistic regression has multiple outcomes. Binary logistic regression, which is applied in this paper, allows researchers "to study how a set of predictor variables is related to a dichotomous response variable" [6]. If we note (X_1, X_2,…, X_n) the set of n explanatory variables, (β_0, β_1,… β_n) the set of n+1 parameters, and Y the dependent variable, the logit model is as follows:

$$\log it(P(Y=1)) = \log it(p) + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...+ \beta_n X_n \quad (1)$$

If we note P(Y=1) = p, the logit function is as follows:

$$\log it(p) = \log \frac{p}{1-p} = \log \frac{P(Y=1/X)}{P(Y=0/X)} = \log(Odds) \quad (2)$$

The model measures the estimated probability of the predicted output, which varies between 0 and 1, and is based on a sigmoid function which has the following form:

$$f(t) = \frac{1}{1+e^{-t}} \quad (3)$$

2.3. Performance measures

To analyze the performance of the binary logistic regression, a number of metrics can be used. One of them is the Hosmer–Lemeshow test [7]. It indicates the existence or not of an important gap between the predicted values and the observed ones [7].

Another important metric used to measure the performance of the logistic regression model is confusion matrix. "Each line of the confusion matrix represents the real or observed class whereas each column represents the predicted class" [5]. "The terms '+' and '-' can be replaced by all the possible outputs of a binary classification problem: true or false, yes or no, etc." [4]. Table 1. below illustrates the confusion matrix. Let us note that True Positive means that the model correctly predicted the value that we defined as Positive. False Positive means that the model has incorrectly predicted the Positive value. False Negatives and True Negatives are interpreted in the same way.

Table 1. Confusion matrix.

		Predicted class	
		+	-
Observed class	+	True Positives	False Positives
	-	False Negatives	True Negatives

From the confusion matrix, important metrics can be deduced: (we note TP for True Positives, FP for False Positives, FN for False Negatives and TN for True Negatives).

- Precision: it reflects the exactness of the positive predictions and has the following formula:

$$precision = \frac{TP}{TP+FP} \quad (4)$$

- Sensitivity: it measures the "true positive rate of the model" [8] and is as follows:

$$Sensitivity = \frac{TP}{TP+FN} \quad (5)$$

- Accuracy: it measures how often the classifier is correct and has the following formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

3. Results and discussion

The aim of this work is to predict whether a subscriber to a car loan is likely to pay it back before termination date or not. Since logistic regression is one of the most commonly used algorithms for solving classification problems,

we apply, in this part, binary logistic regression to our dataset in order to predict the probability of prepayment risk using a number of variables that characterize the subscribers. To estimate our model, we consider a dataset containing 25 601 subscribers to car loans. The variables that describe these individuals and the loan's characteristics are listed below. Table 2. describes these variables and their type:

Table 2. Variables description.

Name of the Variable	Description	Type	Notation
Repayment_Status	The loan repayment status	Categorical	Y
Loan_Amount	The total amount of the loan	Quantitative	X1
Interest_Rate	The rate applied to the loan	Quantitative	X2
Fiscal_Power	The fiscal power of the car bought	Quantitative	X3
Subscribtion_Age	The age of the subscriber at the starting date of the loan	Quantitative	X4
Loan_Duration	The theorical duration of the loan	Quantitative	X5
Income_Level	The income level of the subscriber	Categotical	X6

Our model's dependent variable is Repayment_Status. It is a dichotomous variable that provides information on whether the loan has been paid back at the termination date or fully prepaid. The rest of the explanatory variables are quantitative, except for Income_Level, which is a dichotomous variable that contains two categories: Low Income and High Income.

Since logistic regression is a supervised machine-learning algorithm, it must be trained on a part of the dataset and tested on the rest. Therefore, the application of this algorithm requires splitting the dataset in two categories as follows: a training set, containing 80% of the total number of observations and a test set, containing the rest. Table 3. below shows the application of this split to our dataset.

Table 3. Dataset split.

Dataset	Number of observations	Percentage
Training set	21 334	80%
Test set	4 267	20%
Total	25 601	100%

3.1. Model estimation

Our estimated logit model has the following general formula, where Y represents the variable Repayment_Status and '1' represents prepayment while '0' represents payment at maturity:

$$\log it(\hat{P}(Y=1)) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_6$$

In order to estimate the parameters of the model, we use maximum likelihood estimation and we find the estimated parameters of our model as illustrated in Table 4. below:

Table 4. Model estimation.

Explanatory variables	Estimated parameters (b)	Wald	Sig.	Exp (b)
Loan_Amount	0,04	151,823	0	1,04
Interest_Rate	0,39	85,34	0	1,48
Fiscal_Power	0,001	24,67	0	1,00
Loan_Duration	0,55	1614,842	0	1,73
Income_Level	0,085	44,786	0	1,09
Subscribtion_Age	-0,17	507,587	0	0,84
Constant	-3,023	1137,295	0	0,05

In order to have an idea about whether the logistic model is globally significant and convenient to the researcher's problematic, we often use the likelihood ratio test "that compares between the trivial model and the used model" [7]. In order to test the global significance of the model and to find out if the explanatory variables influence simultaneously the prepayment risk, we use likelihood ratio test. The results found in Table 5. show that the explanatory variables of our model explain the probability of prepayment.

Table 5. Likelihood ratio.

-2log-likelihood	CHI_square	Sig
24407,247	5109,761	0

After validating the binary logistic model, we must test the significance of each explanatory variable. We use the Wald test for that purpose. "The selection procedure is based upon the p-value from the Wald Chi-Square test" [9]. Moreover, in order to find the variables that explain the most the output variable, we use the "Odds Ratio" that has the following formula, where b is the maximum likelihood estimator associated to the variable X.

$$Odds(X) = \frac{P(Y=1/X)}{P(Y=0/X)} = \exp(b) \tag{7}$$

The significance of each explanatory variable is measured through the Wald test [7] and summarized in Table 4. The Wald test shows that all of the variables are significant. Moreover, the sign of the estimated parameters represented in the column b as well as the sign of the odds ratio represented in the column of Exp(b) [7] indicate the way the correspondent explanatory variable is related to the dependent variable. For instance, when the duration of the loan or its interest rate increase, the probability of the loan getting prepaid increases too. However, the opposite happens when the age of the subscriber increases. Therefore, these results represent the effect of each variable on prepayment risk.

Consequently, the logit model used to predict the probability that each subscriber prepays its loan is significant and its attributes significantly explain this prepayment.

3.2. Model evaluation

In order to evaluate the quality of fit of our model, we use a number of metrics such as the confusion matrix and the Homer Test.

The confusion matrix is used to evaluate the performance of our logit model. It represents correctly and incorrectly predicted fitted values and "compares predicted data to observed data" [4]. For our use case, Table 6. below represents the confusion matrix of our model.

Table 6. Confusion matrix of our model.

		Predicted		
		Repayment_Status		
Observed		Prepaid	Paid at term	Overall percentage
Repayment_Status	Prepaid	8 406	1 704	83,15%
	Paid at term	1 537	9 687	86,3%
Overall percentage		9 943	11 391	85%

This confusion matrix is a cross tabulation of the observed values and the predicted ones. Using the results of this matrix, we can calculate the performance measurement metrics and evaluate the performance of the model by calculating the accuracy, the recall (sensitivity) and the precision.

When we apply the formulas (4), (5) and (6) on our data, we find the following results:

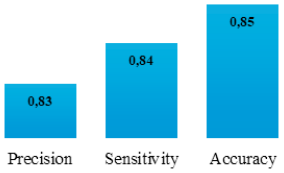

Fig. 2. Precision, Sensitivity and Accuracy

We can see that our model has classified 83% of the data correctly (precision), that the classifier has a 85% accuracy rate and the positive rate of the model is 84%.

The Hosmer–Lemeshow test can also be applied. It indicates the existence or not of an important gap between the predicted values and the observed ones. Table 7. contains the results we found after applying this test to our training data. The results of this test confirm that our model has a good fit:

Table 7. The Hosmer–Lemeshow test.

Chi-square	Sig.
11,489	0,176

We trained our Machine-learning model on 80% of the dataset. The remaining 20% was used to test our trained model. As new subscribers were introduced, the logistic regression model properly predicted the prepayment risk related to them.

4. Conclusion

Subscribers to loans who proceed to prepayment can cause an important loss of profit to banking institutions. Therefore, it is important to have a model that is able to predict whether a customer is likely to pay back the loan before its maturity or not. Moreover, default risk has been the subject of research more than prepayment risk. In this paper, we proposed to apply binary logistic regression as a supervised machine learning classification algorithm, which will represent an aid decision-making tool for banks, which will help to minimize the loss in profit caused by prepayment risk. The logistic model explained the modeled variable that represents the probability of a loan being prepaid before the termination date, by a number of explanatory variables that characterize the loan conditions as well as the subscribers' characteristics. The model's parameters were estimated and their significance tested. Due to the use of performance measurement metrics, we were able to conclude that our model is well fitted and has a good correct classification rate. The logit model proposed has met the purpose of our study.

In terms of future works, we aim to compare the logit model with other classification techniques. Moreover, we aim to conduct a deeper analysis of prepayment by modeling the duration of a loan by survival models such as Kaplan-meier.

References

[1] Abor, Joshua Yindenaba, Agyapoma Gyeke-Dako, Vera Ogeh Fiador, Elikplimi Komla Agbloyor, Mohammed Amidu, and Lord Mensah (2019) "Money and Banking in Africa. Advances in African Economic, Social and Political Development". Springer.

[2] Li, Zhiyong, Ke Li, Xiao Yao, and Qing Wen (2018) "Predicting Prepayment and Default Risks of Unsecured Consumer Loans in Online Lending", *Emerging Markets Finance and Trade*: 118-132.

[3] Batty, Marc, Pirmin Lemberger, Médéric Morel, and Jean-Luc Raffaëlli (2015) "Big Data et Machine Learning manuel du data scientist", Dunod.

[4] Biernat, Eric, and Michel Lutz (2015) "Data science: fondamentaux et études de cas. Machine learning avec Python et R", Eyrolles.

[5] Géron, Aurélien (2017) "Machine Learning avec scikit-learn", Dunod.

[6] Harrell Jr., Frank E. (2015) "Regression Modeling Strategies. With applications to linear models, logistic and ordinal regression, and survival analysis", Springer.

[7] Rakotomalala, Ricco (2017) "Pratique de la régression logistique", Lumière Lyon 2 University.

[8] Guy, Liy (2019) "Application of Machine Learning Algorithms in Predicting Credit Card Default Payment", University of California Los Angeles.

[9] Rezapour, Mahdi, Amirarsalan Mehrara Molan, and Khalid Ksaibati (2019) "Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models", *International Journal of Transportation Science and Technology,* Elsevier.