2020-11

Default avoidance on credit card portfolios using accounting, demographical and exploratory factors: decision making based on machine learning (ML) techniques

Sariannidis, Nikolaos

Springer Verlag

Annals of Operations Research (2020) 294:715–739
https://doi.org/10.1007/s10479-019-03188-0

S.I.: BALCOR-2017

Default avoidance on credit card portfolios using accounting, demographical and exploratory factors: decision making based on machine learning (ML) techniques

Nikolaos Sariannidis[1] · Stelios Papadakis[2] · Alexandros Garefalakis[2] · Christos Lemonakis[2] · Tsioptsia Kyriaki-Argyro[3]

Published online: 15 March 2019

Abstract
Effective and thorough credit-risk management is a key factor for lending institutions, as significant financial losses can arise from the borrowers' default. Consequently, machine learning methods can measure and analyze credit risk objectively when at the same time they face increasingly attention. This study analyzes default payment data from a credit cards' portfolio containing some 30,000 clients from Taiwan with twenty-three attributes and with no missing information. We compare prediction accuracy of seven classification methods used, i.e. KNN, Logistic Regression, Naïve Bayes, Decision Trees, Random Forest, SVC, and Linear SVC. The results indicate that only few out of most of the typical variables used can adequately analyze default characteristics in terms of lending decisions. The results provide effective feedback to credit evaluators, lending institutions and business analysts for in-depth analysis. Also, they mention to the importance of the precautionary borrowing techniques to be used to better understand credit-card borrowers' behavior, along with specific accounting, historical and demographical characteristics.

Keywords Debt · Credit card portfolios · Machine learning (ML) methods · Explanatory factors · Accounting data · Demographic data · Credit history data

1 Introduction

Classification methods allow one to predict the category to which a state of financial condition belongs based on key characteristics. The prediction of credit default is a typical example of applying machine learning techniques. The credit card markets, in recent years, have been growing globally. However, it is a fact that credit card is a high-risk product, which is why credit card interest rates are kept at high levels. It is therefore expected that, as financial institutions seek to maximize their profits, they also strive to limit their bad debts of insolvent customers. At the heart of the banking institutions, their principal

✉ Nikolaos Sariannidis
nsarianni@teiwm.gr

Extended author information available on the last page of the article

concerns—among others—are to adequately manage of the increased credit risk, involves in credit card payments and their use in general.

The high risk of credit card development has highlighted weaknesses identified worldwide in assessing credit quality for potential customers. This study discusses data processing related to the multivariate classification dataset provided in the UCI Machine Learning Repository. The data contains of some 30,000 clients with twenty-three attributes with no missing information. We look for credit card defaults in all cases with the use of accounting, demographic and credit criteria for acquiring clients' credit position. Since credit card industry has significantly grown over the past years, it is very important for analysts and credit institutions to predict clients' creditworthiness.

The core use of credit card has two main categories: (a) as a method of payment, and (b) as a lending method, that cannot replace other types of loans. However, it is a fact that credit card internationally remains a precarious product. Also, the high uncertainty associated with the spread of the "plastic money" market has also highlighted countries' credit rating conditions (e.g., among others: Aha 1992; Frank and Witten 1998; Frank and Hall 2001; Landwehr 2003; Khandani et al. 2010; Dimitras et al. 2017).

The purpose of this work is to identify key factors of clients' default. The analysis of data using computational intelligence methods, and the classification of clients' creditworthiness by using machine learning techniques underscore relevant issues in research literature. Next step in our research is to locate classifiers that provide both accuracy and simplicity in use, so that results can be easily understood and also be used in selecting creditworthy customers.

The rest of the paper is organized as follows. In Sect. 2, we proceed in providing a literature review of relevant research, defining and characterizing the issue of prediction of default rate in credit portfolios. Section 3 presents the main theoretical classification methods used in the study with their characteristics focused on the topic of credit card debt and we pose research questions. We formally describe the dataset information taken and explaining all accounting, demographical and credit-oriented lending criteria used in the study. In Sect. 5 we analyze the credit card portfolio by taken descriptive statistics, while in Sect. 6, we present the comparative results of the study, and examining the classification accuracy in all cases. Section 7 presents our conclusions and future research.

2 Literature review

The financial issue of credit rating, both in terms of credit card issuance and the provision of various types of loans, has been studied extensively. The literature has demonstrated that card holders, especially those carrying high balances, remain highly sensitive to interest-rates. Therefore, there are almost perpetually seeking for lower credit card interest rates to reducing their cost of money. Studies show that relatively few consumers act to alternate bank services despite dissatisfaction with their banks. Also, various methodologies from the field of statistics and decision-making techniques have been successfully implemented. This section contains some of the work that has been done on this issue.

Hand and Henley (1996) make a review of statistical and non-statistical techniques for customers' credit. They refer to the problem of the ability of customers to repay a loan. Before reviewing the different methodologies, the authors present the most common features used in such problems, some of which are: residence time in an address, applicants' annual income, possession or not of credit cards, clients' age, purpose of the loan, etc.

In addition, it is reported that some of the features are categorical data, while others are numeric. The most used methods for such problems, according to this study are: partition analysis, regression, mathematical programming methods, decision trees, smart systems, neural nets and non-parametric methods of normalization. This work concludes that the use of neural nets provides relatively good and stable results to specific credit risk problems.

In addition, Davis et al. (1992) in their work apply various techniques for assessing customer creditworthiness regarding credit cards portfolios. The methods used were the ID3 algorithm, the Multilayer Perceptron Neural Network and neural networks with the use of backpropagation. As a case study, a sample of credit cards was used by the Bank of Scotland, and a decision-making system placed was calibrated. Machine learning methods produced relatively adequate results with respect to the existing system.

In another study, Shi et al. (2005) search for the behavior of credit card holders in relation to their obligations. They propose a quadratic planning method, which takes into account multiple criteria analysis, with a view to classify the behavior of credit card holders. The main reason for using such a technique is the non-linear nature of such problems. The data set has emerged from a large US bank, and a comparison of the results between the proposed methods is made. The numerical results show that the proposed method is a very promising one.

He et al. (2004) research study the problem of classifying credit card users regarding their financial behavior. It is reported that several grading techniques have already been used for this problem, which are based on linear programming for multiple criteria. The contribution of this paper is to introduce a technique combining fuzzy logic with linear programming, with the ultimate scope of identifying trends in the behavior of credit card holders. The training data for this work concerned some 1000 credit card cases from a large US bank, each of which had some 65 characteristics. The solution stemming from the training data was used to predict the behavior of 5000 cases of credit card holders from different states of the USA. The results of the proposed method were compared with those of the multi-criteria linear programming method, the neural network method and the decision trees.

Shen et al. (2007) apply three classification techniques for the problem of identifying fraud conditions in credit cards. Credit card fraud can be divided into two categories: internal and external fraud. In the case of internal fraud, the fraudster attempts to obtain a certain amount of money from the credit card of the legal owner. Instead, in the case of external fraud, there is an attempt of smart purchases to be made using the beneficiary's credit card. The methods used to detect fraud are decision trees, neural networks and regression analysis. The data came from real cases, while the classes of the problem were mainly characterized as fraudulent behavior. In addition, each of the cases had some 18 characteristics. The results show that neural networks and linear regression can adequately detect credit card fraud.

Srinivasan and Kim (1987) apply various classification techniques for business lending. Parametric and non-parametric techniques are used, as well as crisis techniques (e.g., analytical hierarchy). The results show that the use of non-parametric techniques, such as decision trees, can provide very satisfactory classification results.

Lee et al. (2006) resolve the credit rating problem using the CART method and the multivariable adaptive polynomial regression (MARS) technique. The application was made on a bank credit card data set, and the results were compared with traditional techniques such as regression analysis, neural networks and support vector machines. The proposed techniques gave better results in terms of classification. More specifically, with regards to the dataset, some 8000 credit card cases from a Taiwanese bank were used, divided as

follows: 4000 random cases for training sample, 2000 random cases for control sample, and 2000 random cases for validation. Each of the 8000 cases had nine characteristics, i.e.: gender, age, marital status, level of education, occupation, job, annual income, residence and credit limits.

Peng et al. (2004) examine the ability of the multi-criteria linear programming approach to the problem of classifying the behavior of credit card holders in two or more categories using the cross-validation technique as well as clustering techniques. The aim of the work was to examine the stability of the proposed method with and without the aforementioned techniques. The results show that the proposed method is stable, as the correct classification rates with or without the use of cross-validation and clustering techniques are about the same. The data set consists of some 5000 credit card cases from a large US bank. Each case has 113 features (i.e., 38 primary and 65 secondary ones). The 38 primary variables belong to 5 categories: account balance, purchase, payment, cash withdrawal and related variables. In recent years, attempts have been made to use techniques from a particular branch of artificial intelligence techniques, in particular methods based on how operational systems work, for financial classification problems.

Marinakis et al. (2009) use two algorithms that belong to this industry to assess corporate credit risk. More specifically, the algorithm, based on the function of an ant colony and the algorithm used is based on the behavior of a cluster of particles. The application consists of some 1300 data from non-financial corporations in UK, and the classes of the problem are five. The results are encouraging in the performance of the techniques used.

Bhaduri (2009) applies an algorithm, based on how the human immune system works, about the credit rating problem. More specifically, two different datasets are used, from the Australian and German markets. In addition, four different modifications of the algorithm, based on the human immune system, were applied. Comparison with the results of other classification techniques is not clear on the performance of the proposed algorithm. That is, while in one set it produces satisfactory results, in the other its performance falls vertically. One possible cause for this case is that the algorithm based on the human immune system can be considered as a general classification algorithm (i.e., as opposed to the other algorithms that may be adapted to the problem), and this may be an indication for further study and analysis of this technique.

Yeh and Lien (2009) use a dataset of default payments in Taiwan and compared the predictive accuracy of probability of default among six data mining methods. They found that predictive accuracy of the estimated probability of default is more valuable than the binary result of classification for credible or not credible customers. They use a simple linear model to show that artificial neural network is the only one that can accurately estimate the real probability of default of the portfolio used.

In a more resent study Hamori et al. (2018), analyze default payment data and compare prediction accuracy and classification ability of three ensemble-learning methods with those of various neural-network methods. The results focus on the classification ability which is superior to other machine-learning methods including neural networks.

In general terms, research tendency after the year 2011 is placed towards the use of hybrid formulas to solve similar problems, or even the use of more contemporary methodologies from the field of computational intelligence. One reason that explains the global success of these techniques is the increased complexity of the solutions space to specific problems, especially where one takes into account many features or different classes of a problem. In such cases, linear techniques fail to capture the dynamics of the problem, which is achieved by computational intelligence techniques, based on machine learning methods.

3 Methodology

Quite a few machine learning methods can be used for classification of credit cards customers based on explanatory factors regarding accounting, demographical and credit historical records can be used to determine predictive accuracy (e.g., Yeh and Lien 2009; Dimitras et al. 2017; Hamori et al. 2018). These advanced methods underscore a client's ability to repay one debt or to proclaim the default status on next payment with a high accuracy. In this study we make comparison of machine learning methods to achieve most cost-effective prediction for credit card default.

Present approach of implementing a risk control is an attempt to answer the following two research questions:

1. Hypothesis no. 1

How does the probability of default payment vary by categories of different accounting, demographic and credit history factors?

2. Hypothesis no. 2

Which variables are the strongest predictors of default payment?

In this study we use seven methods of classification to answer to the above research questions. More specifically, we use the:

1. K-Nearest Neighbor (KNN) method

KNN is a non-parametric technique, and in its classification, it uses k, which is the number of its nearest neighbors, to classify data to its group. With this measure the distance between all points is calculated and then the k points are found that are closest based on the previously calculated distances. Finally, the class is chosen by the majority of the surrounding points (Cheng et al. 2014). The K is a positive integer and whenever it takes the value of 1, it means that the model is classified to the class of the single nearest neighbor. When KNN method is used for classification problems, the output can be calculated as the class with the highest frequency from the K-most similar instances. KNN method can find very complex patterns but its output is quite challenging to interpret.

2. Logistic regression method

Logistic regression is named for the function used at the core of the method, the logistic function. The logistic function (Makalic and Schmidt 2010), also called the sigmoid function was developed to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that takes values between 0 and 1, but never exactly at those limits.

$$\frac{1}{1 + e^{-value}}$$

where e is the base of the natural logarithms (i.e. Euler's number or the exponential function) and value is the actual numerical value of transformation. In our classification problem we use the Maximum-likelihood estimation, a common learning algorithm to search

for the best coefficients that result in a model that predicts a value very close to 1 for the default class and a value very close to 0 for the non-default class in credit cards portfolio. The procedures search for Beta values that minimize the error in the probabilities predicted (Makalic and Schmidt 2010).

3. Naïve Bayes classifiers method

In classification problems, Naïve Bayes method estimates the probability of a given data point falling in a certain class. Before prediction one should find the parameters for the credit factors individual probability distributions, taking part in the classification problem. Then the method estimates the probability of a given data point by picking the c_i that has the largest probability given the data point's features, as shown in the following formula.

$$y = \arg\max P(c_i) \prod_{j=1}^{n} P(x_j|c_i)$$

This is referred to as the Maximum A Posteriori decision rule and it only used the P(B|A) and P(A) terms, which are the likelihood and prior terms, respectively. In case that the probability P(B|A) is used, the Maximum Likelihood decision rule is taken (Ramoni and Sebastiani 2001).

4. Decision tree method

Decision trees (Quinlan et al. 1998) are representations that classify a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. Some common usages of decision tree models include the variables selection for monitoring the problem (herein defining accuracy of the default status of a credit cards portfolio), assessing the relative importance of variables (i.e. the most important ones that take important part in credit decision) and making the prediction of the default rate, which remains one of the most important usages of decision tree models. Using the tree model derived from historical data, it's easier to predict the result for future records. Key components of a decision tree model are nodes and branches and the most important steps in building a model are splitting, stopping, and pruning the tree (Jenhani and Nahla 2008).

5. Random Forest method

Random Forest trees (Breiman 2001) consist of a set of independent classification trees that work as follows: Initially, a number of trees to grow is selected. In this study we use the number obtained from the type log (M + 1), where M is the number of independent variables. Then a bootstrap sample is selected from the data taken, whereas data does not belong to the bootstrap sample is called "out-of-bag" data. A random tree grows, each node of which is chosen for the best separation between log (M + 1) randomly selected variables, herein the accounting, demographical and credit history data taken from the credit cards portfolio used. The tree grows to the maximum size it can take without being pruned. Then the tree created is used to make the prediction. Random Forest tree classification accuracy is due to minimizing the correlation between classification trees.

6. Support Vector Clustering (SVC) method

In Support Vector Clustering (SVC) algorithm points are mapped from data space to a high dimensional feature space using a Gaussian kernel, where one is looking for the smallest sphere that encloses the image of the data. This sphere taken is mapped back to data space, to form a set of contours that enclose the data points. These contours are taken as cluster boundaries, where points enclosed by each contour are directly related to the same cluster. As the width parameter of the Gaussian kernel is decreased, the number of disconnected contours in data space increases, leading to an increasing number of clusters (Ben-Hur et al. 2001). SVM method advantage is that it is less computationally demanding than kNN, for example, and is easier to interpret but it can identify only a limited set of patterns.

7. Linear Support Vector Clustering (SVC) method

Similar to SVC but it has more flexibility in the choice of penalties and loss functions and can scale better to large numbers of samples, like the one we use in our study with some 30,000 observations from Taiwan. This learning strategy, introduced by Ben-Hur et al. (2001), is a quite powerful method that has already outperformed most other systems in a wide variety of applications. Linear SVM is based on the idea of hyper-plane classifier but with linear separability.

The above methods have been proved to have a broad use and furthermore a success in many financial areas, such as the one of classification of credit portfolios. However, the training time for SVM is at least $O(N^2)$, where N represents the training data set size N, that makes it non-favorable for large data sets. The rest of the methods, i.e. Random Forest Method, Decision Tree Method and KNN methods used have been proposed to enhance machine learning classification success rates towards increasing performance and accuracy.

4 Dataset information and variables used

The dataset taken in this study uses information on a credit card portfolio client in Taiwan derived from a time period between April 2005 to September 2005. It contains customers' default payments on their credit card repayments next month and accounting, demographical, credit factor data along with card holders bill statements.

We use the UCI Machine Learning Repository, available from University College Irvine. The collection has more than 300 datasets, some of which (about 16) are in the "Business" category. Data sets are for education and research in Mechanical Learning. There are sets suitable for cluster categorization, regression, cluster and classification analysis. The site provides information about the data, such as the number of columns and rows, the year of creation, the type of data, jobs for which they are appropriate. In this study, we seek to identify which variables determine the default condition in a credit card portfolio. More specifically, we use an array of twenty-three (23) variables, with a variety of exploratory characteristics as shown in Table 5. We employ a binary variable, default payment (Yes = 1, No = 0), as the response variable.

The description of the variables used follows in Table 1. The dataset contains some 30,000 observations from Taiwan and is characterized by twenty-three (23) explanatory factors that refer to specific customer data regarding accounting, demographical and

Table 1 Variables used

Encoding	Meaning
–	ID: ID of each client
X_1	LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
X_2	SEX: Gender (1 = male, 2 = female)
X_3	EDUCATION: (1 = graduate school, 2 = university, 3 = high school, 4 = others, 5 = unknown, 6 = unknown)
X_4	MARRIAGE: Marital status (dummy variable that takes values of 1 = married, 2 = single, 3 = others)
X_5	AGE: Age of the card holder in years
X_6	PAY_0: Repayment status of the card holder in September 2005, i.e.: − 1 and 0 = duly payments, 1 = payment delay for 1 month, 2 = payment delay for 2 months, 3 = payment delay for 3 months, 4 = payment delay for 4 months, 5 = payment delay for 5 months, 6 = payment delay for 6 months, 7 = payment delay for 7 months, 8 = payment delay for 8 months, 9 = payment delay for 9 months and above
X_7	PAY_2: Repayment status of the card holder in August 2005 (scale same as above)
X_8	PAY_3: Repayment status of the card holder in July 2005 (scale same as above)
X_9	PAY_4: Repayment status of the card holder in June 2005 (scale same as above)
X_{10}	PAY_5: Repayment status of the card holder in May 2005 (scale same as above)
X_{11}	PAY_6: Repayment status of the card holder in April 2005 (scale same as above)
X_{12}	BILL_AMT1: Amount of bill statement of the card holder in September 2005 in NT dollar (*)
X_{13}	BILL_AMT2: Amount of bill statement of the card holder in August 2005 in NT dollar (*)
X_{14}	BILL_AMT3: Amount of bill statement of the card holder in July 2005 in NT dollar (*)
X_{15}	BILL_AMT4: Amount of bill statement of the card holder in June 2005 in NT dollar (*)
X_{16}	BILL_AMT5: Amount of bill statement of the card holder in May 2005 in NT dollar (*)
X_{17}	BILL_AMT6: Amount of bill statement of the card holder in April 2005 in NT dollar (*)
X_{18}	PAY_AMT1: Amount of previous payment of the card holder in September 2005 in NT dollar (*)
X_{19}	PAY_AMT2: Amount of previous payment of the card holder in August 2005 in NT dollar (*)
X_{20}	PAY_AMT3: Amount of previous payment of the card holder in July 2005 in NT dollar (*)
X_{21}	PAY_AMT4: Amount of previous payment of the card holder in June 2005 in NT dollar (*)
X_{22}	PAY_AMT5: Amount of previous payment of the card holder in May 2005 in NT dollar (*)
X_{23}	PAY_AMT6: Amount of previous payment of the card holder in April 2005 in NT dollar (*)
–	Default payment next month: Default payment (1 = yes, 0 = no)

(*) Where New Taiwan Dollar (NT Dollar) is the official currency of the Republic of China used in the Taiwan and its surrounding islands

credit information. We use the amount of the given credit in NT dollars that includes both the individual consumer credit and one's family (supplementary) credit. Also, the gender status is a binary variable, that takes the values of 1 for males and 2 for females.

- *Education* status also takes place in this research which takes the values of 1 for graduate school card holders, 2 for university degree holders, 3 for those with high school education and 4 for the rest of the sample. A client's level of training is a very important assessment component for taking a credit decision. The broad assumption includes conflicting views as to when a client appears to have a higher credibility depending on the level of education, either with a positive or a negative effect on the credit default rate (e.g., Ajay and Shomona 2016; Shomona and Ramani 2011; Watanabe et al. 2011).
- *Marital status* which take values 1 for married customers, 2 for single one and 3 for others (e.g. divorced ones). The marital status of the candidate is taken into account by the awarding officer for the assessment of his financial obligations and for determining his social position and attitude.
- *Age* of sample candidates used takes full integer values [21, 79]. Based on age, it can be determined whether the candidate is in the productive period of his life. Clients with very young ages do not inspire analysts with the necessary confidence in the responsibility of the candidate to fulfill his/her obligations. On the other hand, banks are seeking credit cards for young customers to grow their clientele in order to boost their cross-selling activities. Older candidates, on the contrary, are not in favor of future cooperation with a bank.
- *Credit record of customers' payment* is also used to capture a client creditability status. We tracked the past monthly payment records of the dataset, i.e. from April to September of the status year of 2005 as follows: X_6 is the repayment status in September 2005, X_7 is the repayment status in August 2005,…, and X_{11}, the repayment status of the customer in April 2005. The measurement scale for the repayment status is: -1 and 0 are for duly (regular) payments, where 1 represents payment delay for 1 month, 2 represents payment delay for 2 months,…, 8 represents for payment delays for 8 months and 9 represents for payment delays for 9 months and above. It is apparent that as the days of late payment of the credit card installment or return to the current state of the credit increase, the probability of default over the next month also increases.
- *Amount of bill statement* in New Taiwan Dollar (NT Dollar) is taken for the same time period, where X_{12} represents the amount of bill statement in September 2005, X_{13} represents the amount of bill statement in August 2005,…, and X_{17} represents the amount of bill statement in April 2005 consequently.
- *Amount of previous payment* in NT dollar is also used with variables X_{18} representing the amount for each credit card holder paid in September 2005; X_{19} the amount paid in August 2005, …, and X_{23} the amount paid in April 2005.

To evaluate the effectiveness of the methods, the cross-validation procedure (Stone 1974) was applied. Specifically, the fivefold cross-validation procedure has been used to randomly break all 30,000 data into 5 mutually sub-assemblies of 6000 pieces each. Initially, 4 out of 5 subsets were used to train the system. The remaining subset was used to control the system to record the rate of classification accuracy in the "new" data. This process was repeated four more times, the other subsets that could be used as control samples. As a measure of error in each iteration the percentage resulting from the ratio of the incorrect classifications to the total of the cases to be classified was used. The overall error

rate was the average percentage error of the classifications made in the 5 different control samples. Classification accuracy is obtained if the percentage of the total error meter is deduced from the unit.

5 Descriptive statistics

The following tables analyze descriptive statistics of the sample used in this study. The sample is categorized by gender type in relation to the default events in Table 2.

Table 3 depicts that the most defaulted creditors are those who hold a University degree (i.e. 3300 observations). However, the highest percentage per education class, considering the credit defaults in relation to the total of observations, comes from High school graduates, followed by University graduates (24%), Postgraduate students (19%) and other categories (unknown 14%).

Single in marital status clients portray the lowest relative percentage of default rate in the portfolio examined (Table 4), regarding the marital status and the credit attitude of the

Table 2 Gender and credit attitude

Gender	Duly payments and delinquencies (no default)	Defaults	Sum
Men	9015	2873	11,888
Women	14,349	3763	18,112
Total sum	23,364	6636	30,000

Table 3 Education and credit attitude

Categories	(1) Duly payments and delinquencies (no default)	(2) Defaults	(3) Sum	(4) % of default rate [(4) = (2)/(3)] defaults/sum
1 = graduate school	8549	2036	10,585	19
2 = university degree	10,700	3330	14,030	24
3 = high school	3680	1237	4917	25
4 = others	116	7	123	6
5, 6 = unknown	319	26	345	8
Total sum	23,364	6636	30,000	22

Table 4 Marital status and credit attitude

Marital status/categories	(1) Duly payments and delinquencies (no default)	(2) Defaults	(3) Sum	(4) % of default rate [(4) = (2)/(3)] defaults/sum
1 = married	10,453	3206	13,659	23
2 = single	12,623	3341	15,964	21
3 = others	288	89	379	24
Total sum	23,364	6636	30,000	22

Table 5 Age range and credit attitude

Age range	(1) Duly payments and delinquencies (no default)	(2) Defaults	(3) Sum	(4) % of default rate [(4) = (2)/(3)] defaults/sum
20–25	2839	1032	3871	27
26–30	5703	1439	7142	20
31–35	4670	1126	5796	19
36–40	3854	1063	4917	22
41–45	2807	798	3605	22
46–50	1799	601	2400	25
51–55	1072	353	1425	25
56–60	**421**	**151**	**572**	**26**
61–65	**136**	**50**	**186**	**27**
66–70	53	18	71	25
71–80	**10**	**5**	**15**	**33**
Total sum	**23,364**	**6636**	**30,000**	22

With bold letters the highest levels of relative default rates shown from the sample of the study is shown

customers respectively. Other categories (e.g., divorced in marital status clients) show the highest rate of default that is adding up to a 24% overall, with an average reaching at 22%.

Clients' ages with the highest incidence of default status are placed in the range of 71–80, 61–65 and 56–60 respectively, which is a fact firmly expected on the basis of empirical data, showing that credit lending and debt in general becomes almost a prohibitive condition for clients with ages over 70 years old (Table 5).

The analysis of Table 6 shows that for September's 2005 delays in credit cards, the highest relative percentage change is placed in the 2-month payment delay (in absolute terms), with 69% relative to the total sample observations. It is important to note at this point that for clients' delays in installment payments for 2 months or more, default rate tend to become much higher than dully payments; this pose a feature clearly shows that the default payment situation is directly related to the empirical notice, a parameter that arises from late payment of a debt; and above all that clients presenting debts of more than 90 days (i.e.: 3 months), are considered to be practically as defaulted ones.

6 Comparative results

Knowing core characteristics of credit card clients is the first step for credit institutions to set out appropriate strategies or for policy makers to implement borrowers' protection plans. Empirical results are consistent with a general perspective that credit card users are more likely to be married, better educated and property owners. Literature review shows that consumers' borrowing behavior is highly affected by a substantial number of external conditions. For example, the increase in credit limit is followed by an immediate and significant rise in the credit card debt which increases at the same time the default rate in credit cards portfolios. The effect of higher credit limits, indicate that consumers with liquidity constraints tend to increase their debt, which ultimately poses a greater volume of credit claims in question.

Table 6 Repayment status of the card holder (September 2005)

Status	(1) Duly payments and delinquencies (no default)	(2) Defaults	(3) Sum	(4) % of default rate [(4) = (2)/(3)] defaults/sum
Duly payments = -2, -1, 0	19,975	3207	23,182	14
Payment delay for 1 month = 1	2436	1252	3688	34
Payment delay for 2 months = 2	823	1844	2667	69
Payment delay for 3 months = 3	78	244	322	76
Payment delay for 4 months = 4	24	52	76	68
Payment delay for 5 months = 5	13	13	26	50
Payment delay for 6 months = 6	5	6	11	55
Payment delay for 7 months = 7	2	7	9	78
Payment delay for 8 months = 8	8	11	19	58
Total sum	23,364	6636	30,000	22

In our study, the comparative results of the classification accuracy of the different methods applied to the credit rating problem are presented in the following Table 7, where a fivefold cross validation is used with forward inclusion as an attribute selection method.

The dataset used is imbalanced with a 77.8% of instances being NO (0) meaning that a model that always says NO (reject) has a success rate of 77.8% or a 22.12% error factor. We can see from Table 7 that the best results in the initial settings were obtained in the fivefold cross validation providing an overall accuracy of 81.65% (all parameters included) in the SVC methodology with the Radial Basis Function (RBF) of the kernel in use and the Gamma parameter equals to 3. In almost all cases, the methods studied are giving success rates at the same level. Though small differences in success rate do occur whenever the important attributes determined per method are considered, and all factors used, where we see accuracy in prediction moving from 80 to 82.65%, which shows a satisfactory predictive capacity of methods. However, when all the interpretive variables are taken into account, there is a significant deviation, mainly in the Naïve Bayes and Decision trees methodologies with a percentage reduction of 8–10% compared to the models that the most important variables are included. This result is in parallel with the results of other relevant studies (e.g., Neema and Soibam 2017; Krichene 2017).

Responding to the research questions raised in the part of the methodology, identify the following (in combination with Table 8 see also results in "Appendix" section): the most important predictor in all the seven methods used in the study is the X_6 which represents the repayment status of the card holder in September 2005. By all means this attribute is the most important one also in empirical studies as well as in real world (see also Hamori et al. 2018; Neema and Soibam 2017; Dimitras et al. 2017). Keeping up to date on credit info remains an

Table 7 Comparative results

Model	Attributes	Cvsr (cross validation success rate)	Cver (cross validation error rate)
Knn (k = 5, p = 2)	x_6	81.28	18.72
	All	79.36	20.64
Logistic regression	x_6	81.96	18.04
	All	80.97	19.02
Naïve Bayes	x_2, x_3, x_5, x_6	82.02	17.98
	All	70.94	29.06
Decision tree	x_6, x_7, x_{10}	82.02	17.98
	All	72.68	27.32
Random forest	x_3, x_6, x_{11}	82.04	17.96
	All	80.85	19.14
Linear SVC	$x_3, x_4, x_6, x_7, x_8, x_{12}, x_{13}$	80.24	19.76
	All	80.17	19.82
SVC	$x_3, x_6, x_7, x_{10}, x_{11}, x_{14}$	82.21	17.79
	All	81.65	18.35

Cross-validation avoids overlapping test sets. First step: data is split into 5 subsets of equal size. Second step: each subset in turn is used for testing and the remainder for training. This is called fivefold cross-validation. The subsets are stratified before the cross-validation performed. The error estimates are averaged to yield an overall error estimate

Table 8 Variables of key importance

Methods	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{17}	X_{22}
Decision tree classifier		**X**			**X**	**X**		X						
Gaussian NB classifier	X	**X**		X	**X**									X
K-nearest neighbor classifier					**X**									
Linear SVC		**X**	X		**X**	**X**	X			X	X			
CVSR					**X**									
Random forest classifier		**X**			**X**				X					
SVC		**X**			**X**	**X**		X	X	X		X	X	

Only key attributes are portrayed herein, those who contributing to increased excess rates per method. Therefore, factors X_1, X_9, X_{10}, X_{15}, X_{16}, X_{18}, X_{19}, X_{20}, X_{21}, and X_{23} are excluded from the above table. Highlighted columns X_6, X_3 and X_7 are of key importance respectively for the analysis made

important process which even today ensures that a debt position is adequately monitored and follows a procedure for providing an adequate capital buffer to support banking institution capital adequacy levels to cover possible losses from default.

Contrary to what was initially believed factors considered regarding the Amount of bill statements and Amount of previous payments appear to be less important.

Another important element observed in the analysis is related to the education variable (significant in five out of seven methodologies examined-see Table 8). From the analysis of the sample in Table 2 it is understood that the most creditworthy borrowers in terms of relative quota are those with postgraduate studies, showing that in general the development of skills through education is interpreted as indicating a higher impact by the borrowers in the amount and the necessity of credit line they seek to cover up their needs. Those with postgraduate degrees in education seem to be more conservative in credit lending than those with a University degree. On the other hand, high school graduates are the ones who proportionally show the highest default rate in credit portfolio examined, indicating that this category of borrowers may require a better-quality monitoring for credit assessment.

In addition, the X_7 factor (i.e., Repayment status of the card holder in August 2005) reveals the customer's debt status in a period of 1 month before credit evaluation. It basically proclaims a red flag signal indicating possible default condition during the monitoring process for Credit Institutions.

7 Conclusions and future research

Personal credit with the use of credit cards is systematically held by most consumers in developed countries. This study seeks to pose key characteristics for card holders to generally behaving rationally to maximize their own utility. However, some credit cards clients are still shown to misuse their credit cards, and sometimes suffer exploitation by the credit institutions. The major contribution of the paper lays in introducing key customers' aspects, such as financial data, due payments, and other operational characteristics that place greater emphasis on characterizing them in terms of credibility. Several machine learning algorithms were applied to a credit portfolio for the months of April to September 2005, which contained customer credit card data, and models were used to assess the creditworthiness of these customers. The accuracy of the resulting models ranges from about 70% to 82.6%. Therefore, their accuracy could be considered satisfactory and could therefore be used by financial institutions or credit card companies to categorize prospective customers according to their solvency conditions during the approval process and with the use of less information, instead of treating a great deal of accounting, demographic and credit information.

In particular, there were 30,000 credit card cases, of which some 23,364 of them did not show default condition (i.e. normally payed or with little delays) and some 6636 cases characterized as default conditions, as determined in September 2005 customers' conditions. The applications contained customer data e.g. *Amount of the given credit, Gender, Educational and Marital status, Age of the client, Repayment status in the time period of 6 months, and the Amount of the statement of the cardholders for the same period*, which was used to determine the quality of the provision of clients' default conditions.

The accuracy level achieved can be considered as quite satisfactory, where in the majority of cases reached over 80%; therefore, the proposed framework could be used for

a thorough understanding in credit cards market. Some restrictions are set regarding the results stability of the classification methods used herein, where in some cases the success rate alternating through different credit portfolios. These methods can be further developed by analysts, financial institutions or credit card companies to facilitate classification process for prospective customers according to their preferred characteristics.

The results of the study show that the repayment status of the cardholder 1 month before the statutory period (herein the September 2005), coupled with the client's educational status and the systematic monitoring of cases of indebted borrowers in previous months (i.e. clients' credit record), pose key credit factors; this remarks are also in consistency with other relevant studies (e.g., Hamori et al. 2018; Neema and Soibam 2017; Dimitras et al. 2017). Close monitoring of the above factors in a credit card portfolio reveals that a thorough organized framework should be followed by respective credit institutions to reinforce regularly installments of their card holders. Therefore, the lending decision-maker should take the proposed characteristics further into consideration beforehand and during the credit card acceptance process.

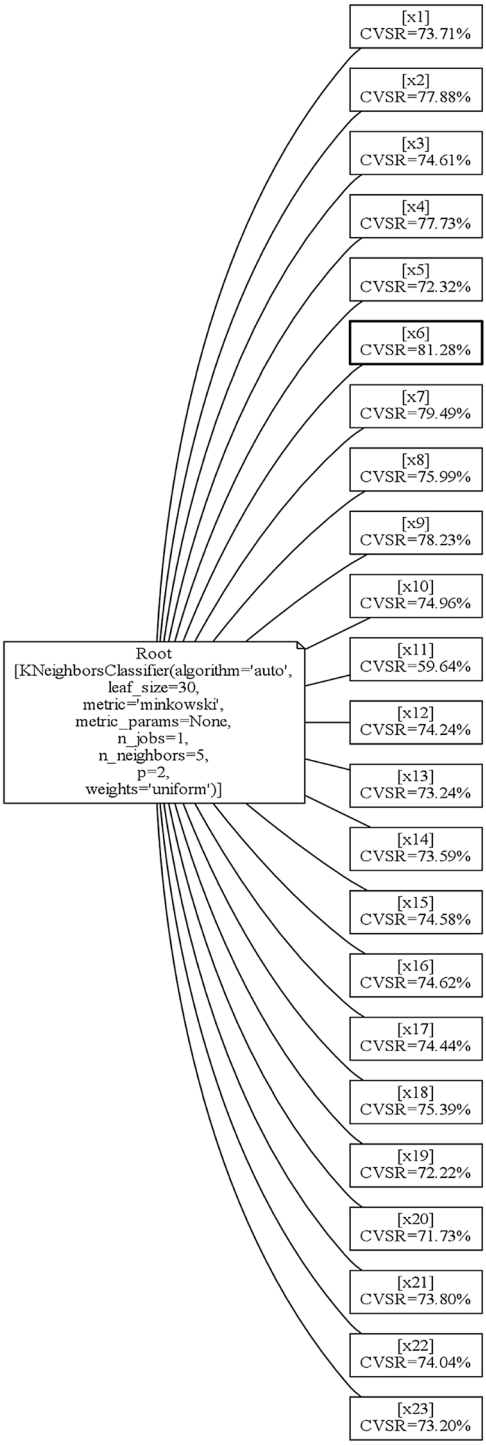
A future consideration, though, should take into account research in the field of other classification methods, encountering more restrictions posed by regulations on the card issuers globally. Also, in some cases (i.e. Naïve Bayes and Decision tree classifications) high terms of error rates in validation of Default clients' status, (e.g. 29% and 27% respectively in this study), when all explanatory variables are included needs to be further examined. These models are characterized by a relative volatility in success rates of forecasting given and therefore it is not likely to be used for credit cards portfolio reviews. It would also be helpful for credit institutions to promote better framework of disclosure information to reduce complexity of the contracts and enhance consumers' understanding, by reducing the amount of irrelevant and partly useless information that does not add anything special to the credit decision and providing a more compact customers' position that summarizes key information.

In the future, financial institutions and credit analysts could even more benefit from the development of machine learning based techniques (e.g. SVC, Random Forest, etc.) identifying more accurately the credit risk groups to which their customers belong to, based on quantitative and operational data. Categorizing clients' characteristics and allocating them into different credit risk groups assists in better understanding and monitoring of banks' loan portfolios and in pursuing its credit policy effectively.

Appendix: Results of classification methods used

See Figs. 1, 2, 3, 4, 5, 6 and 7.

Fig. 1 KNN classification



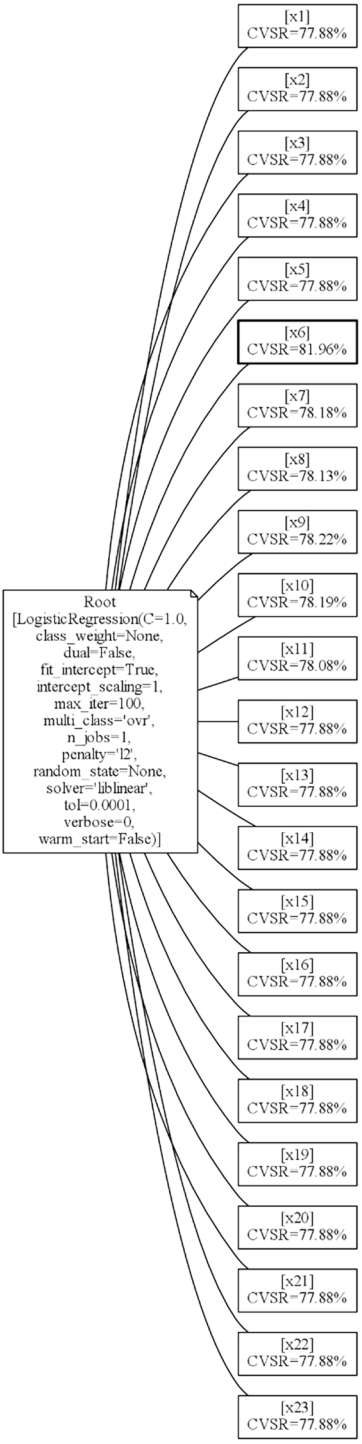


Fig. 2 Logistic regression classification

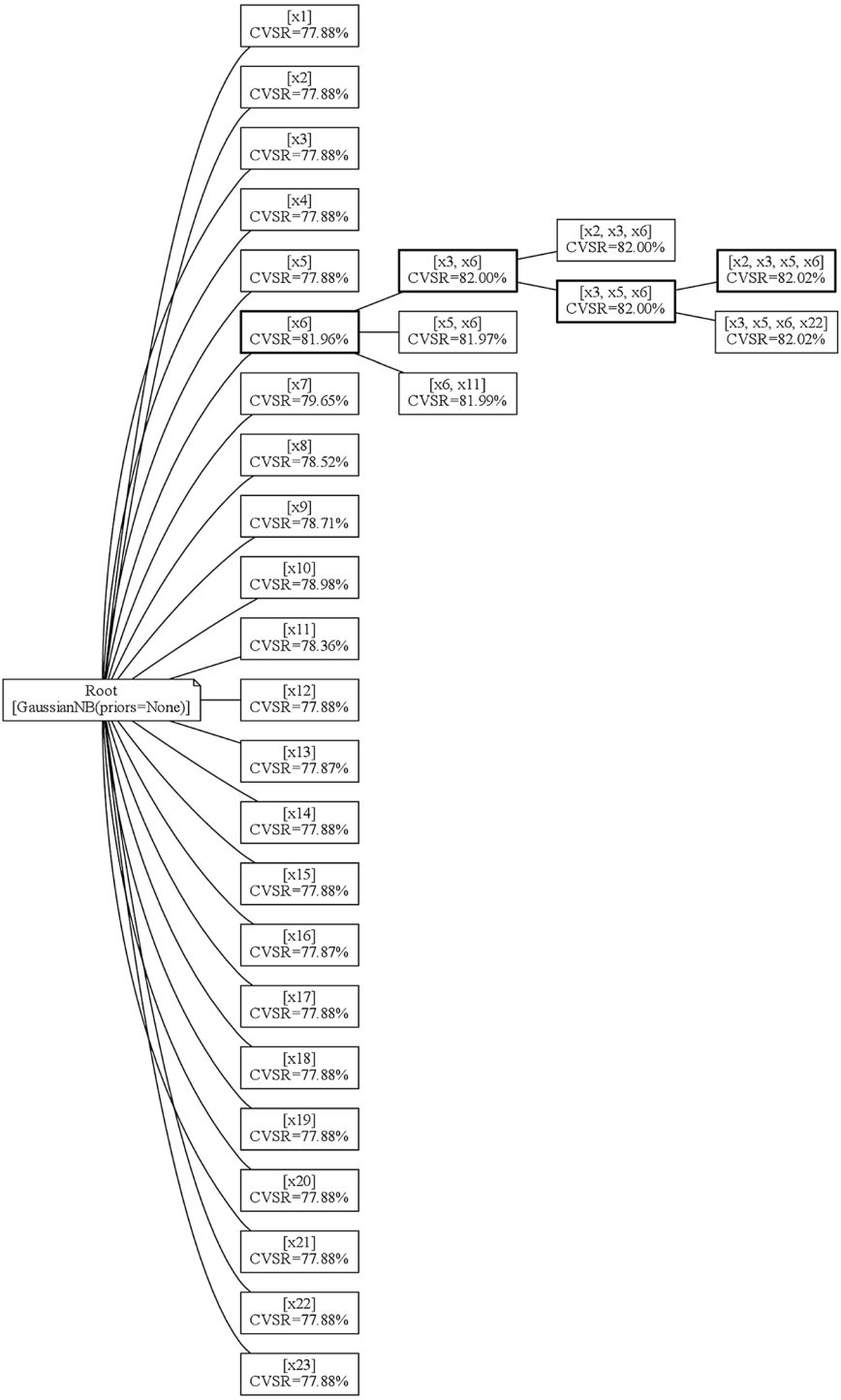


Fig. 3 Naive Bayes classification

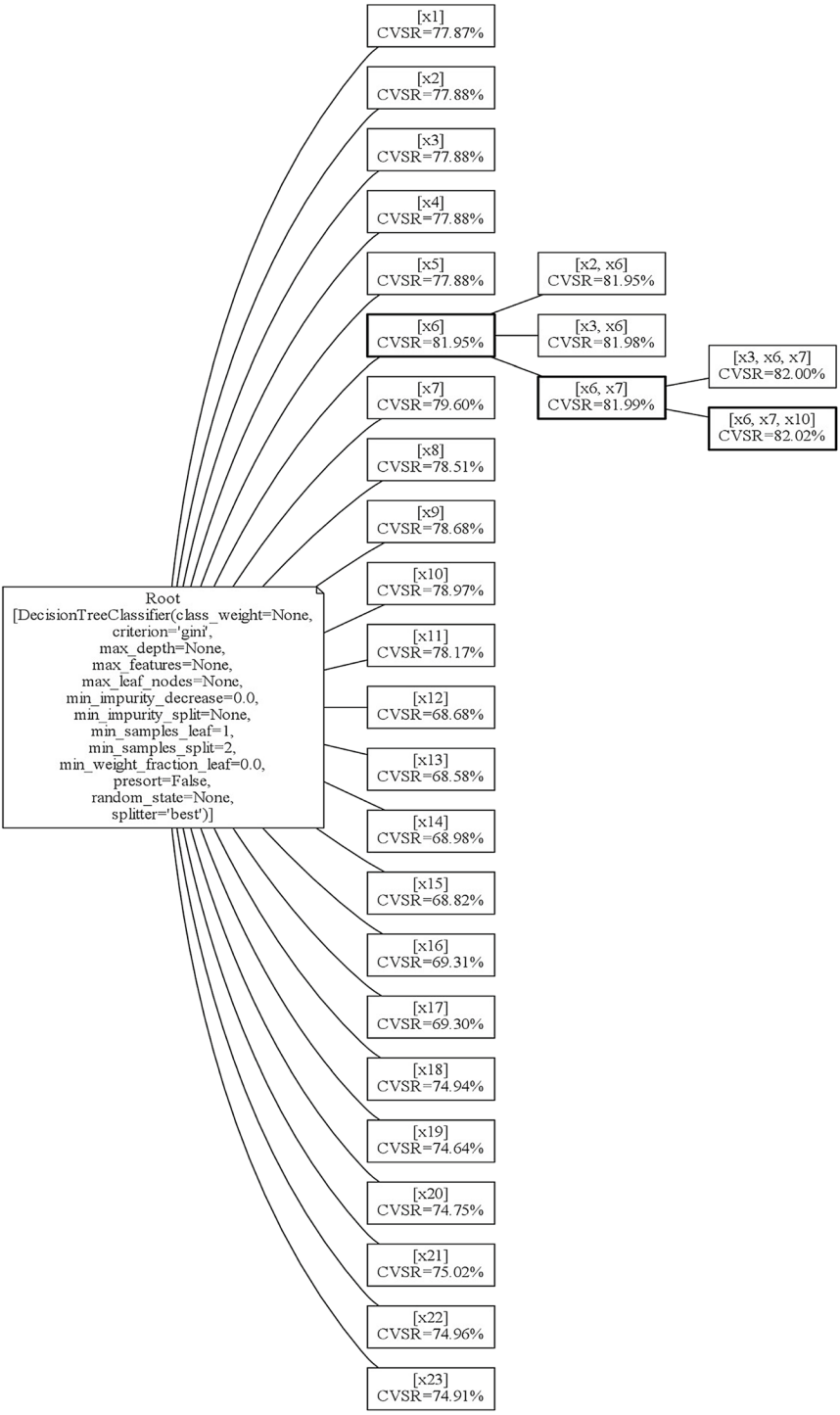


Fig. 4 Decision tree classification

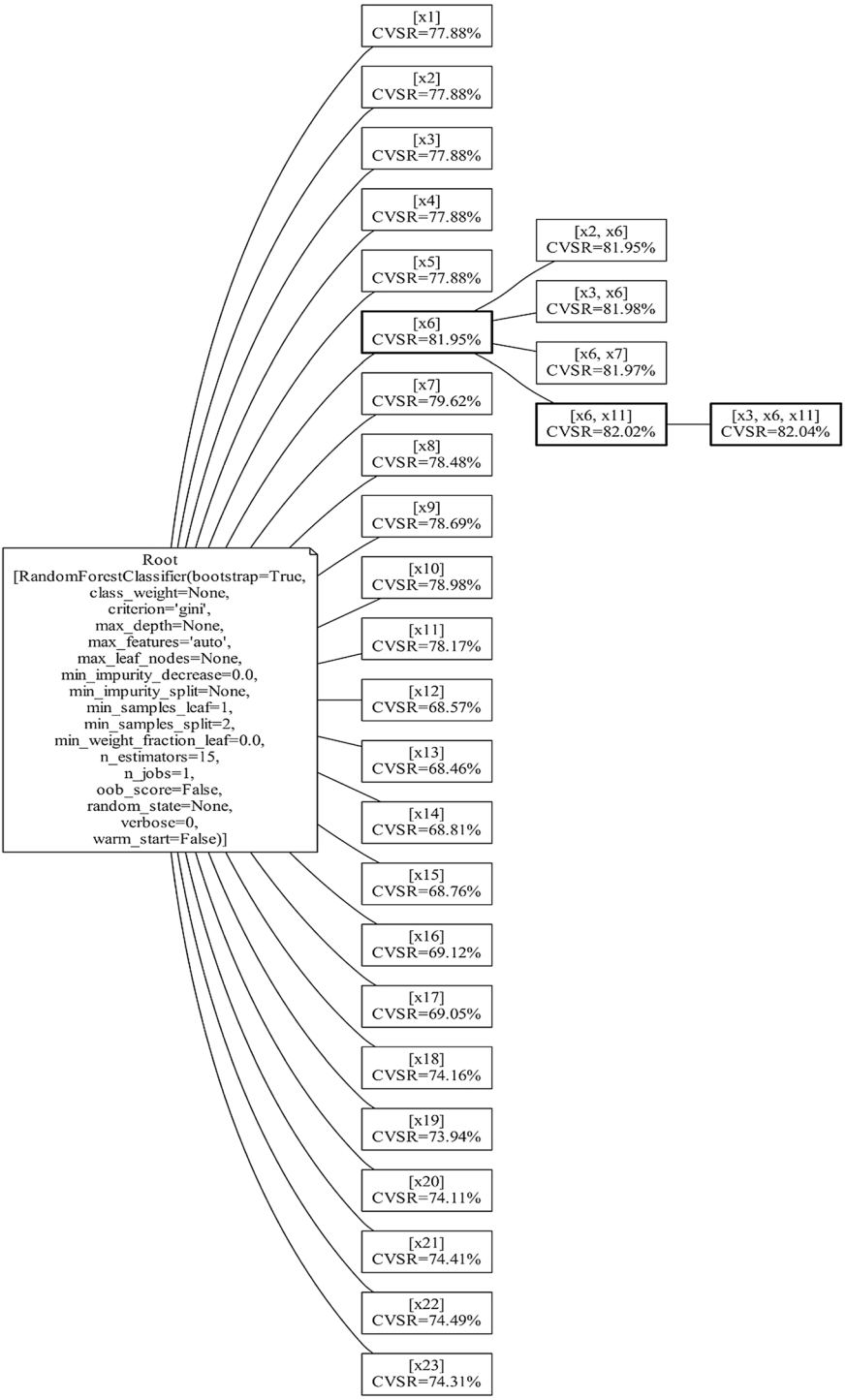


Fig. 5 Random forest classification

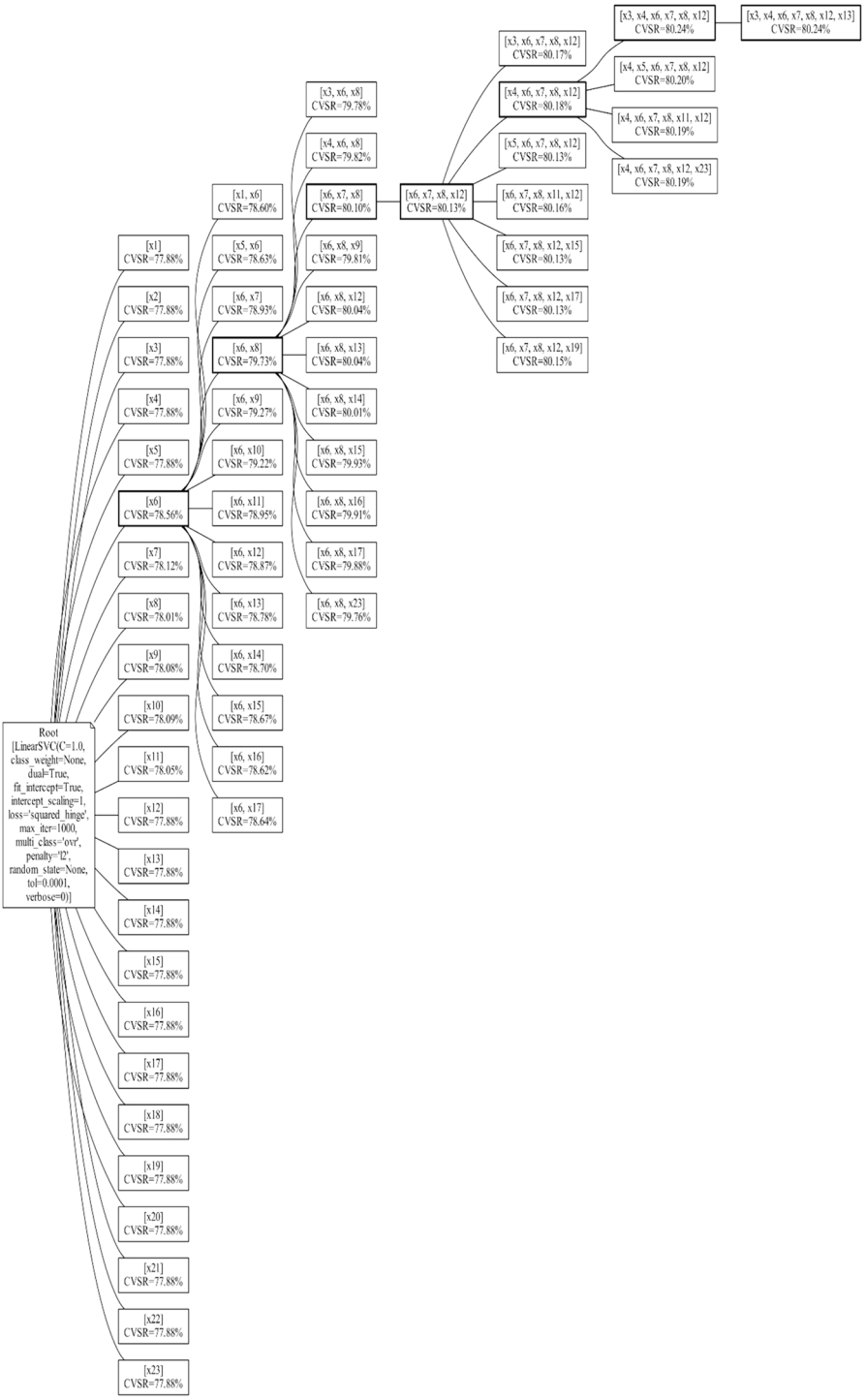


Fig. 6 Linear support vector classification

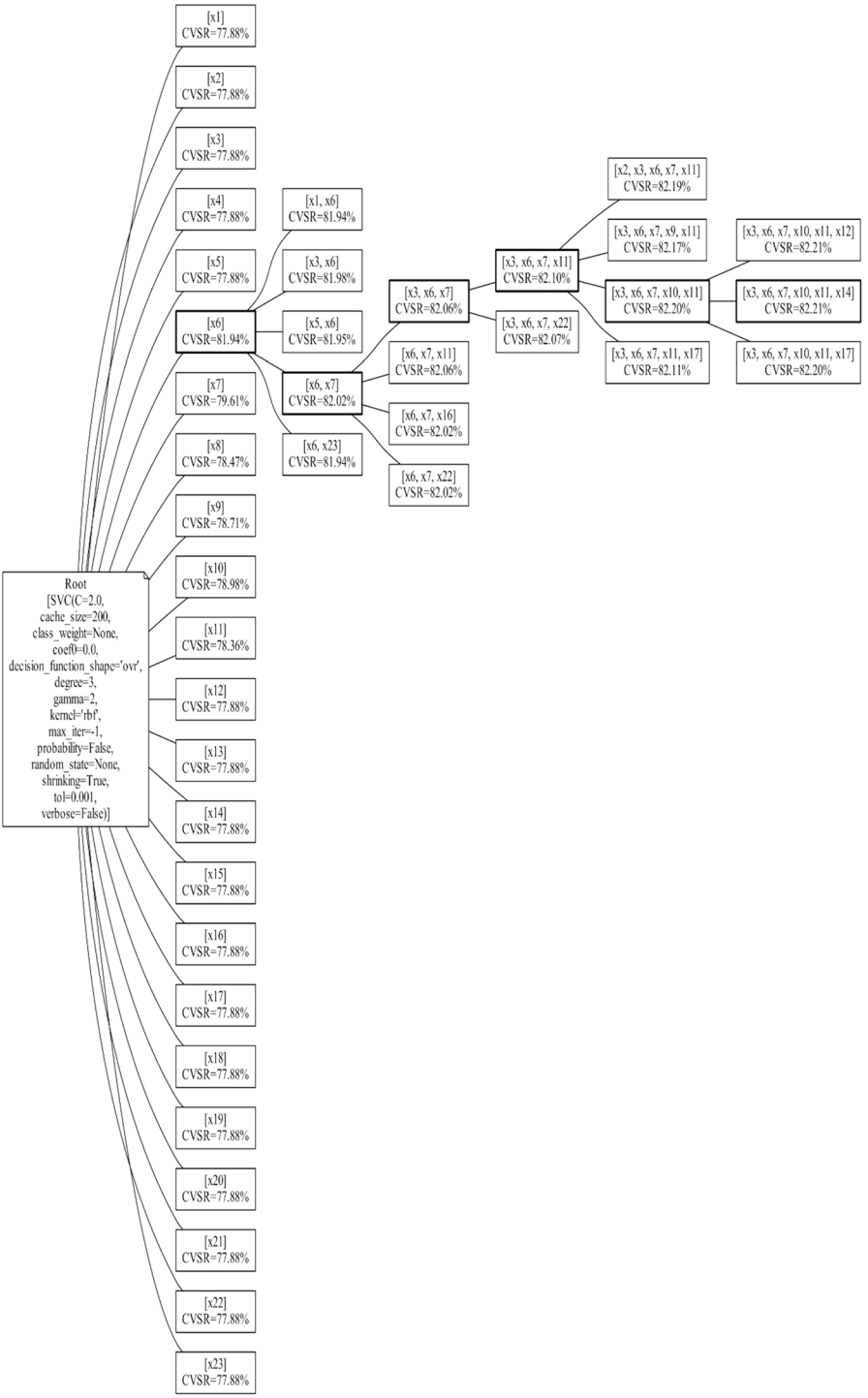


Fig. 7 C-support vector classification

Acknowledgements The current publication is based on the following dataset: Lichman (Lichman 2013). We would also like to thank the Laboratory of Artificial Intelligence Systems and Computer Architectures of the Technological Educational Institute of Crete for providing the computer power to complete extensive experimental results for the needs of this work

Data availability The data set is based on the publicly available credit card default data set from the UCI Machine Learning Repository. Details are here: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

References

Aha, D. (1992). Tolerating noisy, irrelevant, and novel attributes in instance-based learning algorithms. *International Journal of Man–Machine Studies, 36*(2), 267–287.

Ajay, V., & Shomona, G. J. (2016). Prediction of credit-card defaulters: a comparative study on performance of classifiers. *International Journal of Computer Applications (0975–8887), 145*(7), 36–41.

Ben-Hur, A., Horn, D., Siegelmann, H. T., & Vapnik, V. (2001). Support vector clustering. *Journal of Machine Learning Research, 2,* 125–137.

Bhaduri, A. (2009). Credit scoring using artificial immune system algorithms: a comparative study. In *Proceedings of the world congress on nature and biologically inspired computing NaBIC2009, Coimbatore* (pp. 1540–1543).

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

Cheng D., Zhang S., Deng Z., Zhu Y., & Zong M. (2014). *k*NN algorithm with data-driven *k* value. In: Luo X., Yu J. X., & Li Z. (Eds.), *Advanced data mining and applications*. ADMA 2014. Lecture Notes in Computer Science (Vol. 8933). Berlin: Springer.

Davis, R. H., Edelman, D. B., & Gammerman, A. J. (1992). Machine-learning algorithms for credit-card applications. *Journal of Management Mathematics, 4*(1), 43–51.

Dimitras, A., Papadakis, S., & Garefalakis, A. (2017). Evaluation of empirical attributes for credit risk forecasting from numerical data. *Investment Management and Financial Innovations, 14*(1), 9–18. https://doi.org/10.21511/imfi.14(1).2017.01.

Frank, E., & Witten, I. H. (1998). Generating accurate rule sets without global optimization. In J. Shavlik (Ed.), *Proceedings of the fifteenth international conference on machine learning, Madison, WI*. San Francisco: Morgan Kaufmann (pp. 144–151).

Frank, E., & Hall, M. (2001). A simple approach to ordinal classification. In L. de Raedt, & P. A. Flach (Eds.), *Proceedings of the twelfth European conference on machine learning, Freiburg, Germany*. Berlin: Springer (pp. 145–156).

Hamori, S., Kawai, M., Kume, T., Murakami, Y., & Watanabe, Y. (2018). Ensemble learning or deep learning? Application to default risk analysis. *Journal of Risk and Financial Management, 11*(1), 12. https://doi.org/10.3390/jrfm11010012.

Hand, D. J., & Henley, W. E. (1996). A k-nearest-neighbour classifier for assessing consumer credit risk. *The Statistician, 45*(1), 77–95.

He, J., Liu, X., Shi, Y., Xu, W., & Yan, N. (2004). Classifications of credit cardholder behavior by using fuzzy linear programming. *International Journal of Information Technology and Decision Making, 3*(4), 633–650.

Jenhani, I., Nahla, B. A., & Ziedm, E. (2008). Decision trees as possibilistic classifiers (Special Section on Choquet Integration in honor of Gustave Choquet (1915–2006) and Special Section on Nonmonotonic and Uncertain Reasoning). *International Journal of Approximate Reasoning, 48*(3), 784–807.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit risk models via machine-learning Algorithms. AFA 2011 Denver Meetings Paper. https://doi.org/10.2139/ssrn.1568864.

Krichene, A. (2017). Using a naive Bayesian classifier methodology for loan risk assessment evidence from a Tunisian commercial bank. *Journal of Economics, Finance and Administrative Science, 22*(42), 3–24.

Landwehr, N., Hall, M., & Frank, E. (2003). Logistic model trees. In N. Lavrac, D. Gamberger, L. Todorovski, & H. Blockeel (Eds.), *Proceedings of the fourteenth European conference on machine learning, Cavtat-Dubrovnik, Croatia*. Berlin: Springer (pp. 241–252).

Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics and Data Analysis, 50,* 1113–1130.

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine: University of California, School of Information and Computer Science. The original dataset can be found

at the UCI Machine Learning Repository, i.e. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Makalic, E., & Schmidt, D. F. (2010). Review of modern logistic regression methods with application to small and medium sample size problems. In Li, J. (Eds.), *AI 2010: advances in artificial intelligence. AI 2010*. Lecture Notes in Computer Science, (Vol. 6464). Berlin: Springer.

Marinakis, Y., Marinaki, M., Doumpos, M., & Zopounidis, C. (2009). Ant colony and particle swarm optimization for financial classification problems. *Expert Systems with Applications, 36,* 10604–10611.

Neema, S., & Soibam, B. (2017). The comparison of machine learning methods to achieve most cost-effective prediction for credit card default. *Journal of Management Science and Business Intelligence, 2*(2), 36–41.

Peng, Y., Kou, G., Chen, Z., & Shi, Y. (2004). *Cross-validation and ensemble analyses on multiple-criteria linear programming classification for credit cardholder behavior*, Lecture Notes in Computer Science, ICCS 2004 (Vol. 3039, pp. 931–939).

Quinlan, J., Rajendra, G., & Castro, D. (1998). Bank collateralised loan obligations: From 0 to 60 in less than 2 years? Merrill Lynch, Global Securities Research & Economics Group, March.

Ramoni, M., & Sebastiani, P. (2001). Robust Bayes classifiers. *Artificial Intelligence, 125*(1–2), 209–226.

Shen, A., Tong, R., & Deng, Y. (2007). Application of classification models on credit card fraud detection. In *Proceedings of the international conference on service systems and service management, Chengdu* (pp. 1–4).

Shi, Y., Peng, Y., Kou, G., & Chen, Z. (2005). Classifying credit card accounts for business intelligence and decision making: A multiple-criteria quadratic programming approach. *International Journal of Information Technology and Decision Making, 4*(4), 581–599.

Shomona, J. G., & Ramani, R. G. (2011). Discovery of knowledge patterns in clinical data through data mining algorithms: Multi-class categorization of breast tissue data. *International Journal of Computer Applications, 32*(7), 46–53.

Srinivasan, V., & Kim, Y. H. (1987). Credit granting: A comparative analysis of classification procedures. *The Journal of Finance, 42*(3), 665–681.

Stone, M. (1974). Cross-validation choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B, 36,* 111–147.

Watanabe, C. Y. V., Ribeiro, M. X., Traina, C., & Traina, A. J. M. (2011). SACMiner: A new classification method based on statistical association rules to mine medical images. In: J. Filipe, & J. Cordeiro (Eds.), *Enterprise information systems. ICEIS 2010*. Lecture Notes in Business Information Processing (Vol. 73). Berlin: Springer.

Yeh, I.-C., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2, Part 1), 2473–2480.

Affiliations

Nikolaos Sariannidis[1] · Stelios Papadakis[2] · Alexandros Garefalakis[2] · Christos Lemonakis[2] · Tsioptsia Kyriaki-Argyro[3]

Stelios Papadakis
spap@staff.teicrete.gr

Alexandros Garefalakis
alex_garefalakis@yahoo.gr

Christos Lemonakis
lemonakis.christos@gmail.com

Tsioptsia Kyriaki-Argyro
kellytsioptsia795@gmail.com

[1] Department of Finance and Accounting, Western Macedonia University of Applied Sciences, Kozani, Greece

[2] Department of Business Administration, Technological Educational Institute of Crete, Agios Nikolaos Branch, Heraklion, Crete, Greece

[3] Department of Accounting and Finance, Western Macedonia University of Applied Sciences, Kozani, Greece

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature").

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com