IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

CreditWorthines Prediction using Logistic regression: A Machine Learning Approach

K.C.Bhanu[1], Dr.P.Uma Maheswari Devi[2]

[1]Research Scholar, Department of Commerce and Management Studies
Adikavi Nannayya University ,Rajahmundry

[2]Associate Professor , Department of Commerce and Management Studies
Adikavi Nannayya University ,Rajahmundry

bhanu1605@gmail.com,umadevi_4@yahoo.com

ABSTRACT:

Before making a loan to a person, the organisation should examine their creditworthiness to reduce the possibility of risk to their credit. Banks, credit card firms, insurance providers, property managers, governments, and mortgage lenders are some examples of these organisations. A person's three-digit credit score is all it takes to tell lenders whether they can repay a loan within a certain period of time. The better the borrower seems to potential lenders, the higher the credit score. A person's credit history, including the number of open accounts, overall debt levels, payment history, and other characteristics, is used to calculate their credit score. The credit score of an individual can be predicted using a variety of machine learning algorithms. However, due to its desirable qualities, including clarity and robustness, logistic regression is thought to be the most often used model of all. In this study, we employ logistic regression to build a credit-scoring model that can determine whether or not a consumer is reliable based on his credit score.

KEYWORDS:

Credit Score, Classification, Logistic Regression, Binary Logistic Regression, Credit worthiness.

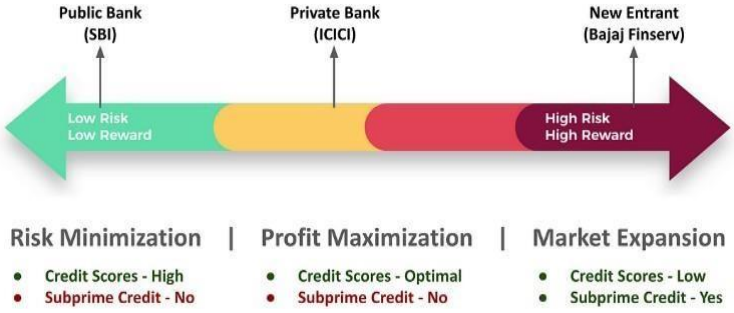
INTRODUCTION:

At some point in our lives, we should all have encountered the term "credit scoring." To determine a borrower's credit score, banks and other financial institutions mostly use statistical analysis. The computation of this statistical three-digit figure, known as a credit score, takes into account the borrower's repayment history, the number of prior credit inquiries conducted, the number of current credit cards or loans, etc. Banks and other financial organisations use this figure as a proxy to assess a loan applicant's creditworthiness. Simply said, banks base their lending decisions on this estimated amount. This decision-making issue is complicated for a bank or any other financial organisation due to a number of considerations. No matter how big or little the bank is, not all debtors would have a substantial credit history to have a decent credit score, or even any credit score. A borrower with a high credit score might not go to a small lender, complicating the decision-making process.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

Prior to the adoption of formal procedures in banking, judgements were made on a judgment-based basis: the bank manager would evaluate a potential applicant's creditworthiness based on his or her personal acquaintance with the applicant. This had a number of flaws, including that it was subjective with all the risks of irrational personal prejudice that implies, unreliable (it can change from day to day with the bank manager's mood), not replicable (another manager may make a different decision, and the reasoning behind the decisions may not be reproducible), hard to teach, incapable of handling large numbers of applicants, and, in general.



Banks, financial institutions, and any other institutions are positioned differently on the risk spectrum and have distinct business strategies depending on the scope of operation, as we can see here.

1. A public bank towards the left of this spectrum would always strive to reduce business risk by only selecting loan applicants with high credit scores.

2. A private bank would aim to maximise its profit, which necessitates that it determines its ideal tolerance levels for credit scores.

3. However, a new entrant that falls to the right of the risk spectrum would be forced to take into account any candidate that knocks on their door, regardless of whether they had a good credit score or none at all.

OBJECTIVES:

For more than 50 years, credit scoring has been utilised as a technique of evaluating credit. Credit cards saw the first successful application of credit scoring. The first retail credit scoring model for credit cards in the US, according to Anderson (2007), was put forth around 1941 and was based on the following criteria for rating credit card applications:

- The position held by the candidate
- The period of time you've held your current position.
- The period of time spent at the current address of residence
- Provided information about bank accounts and life insurance plans
- Gender
- Amount of the instalment due each month

The decision time was demanded to be shortened by the growth in US credit card business. In order to aid in consumer credit evaluation, Fair, Isaac & Co. (FICO) was founded in 1956,

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

and in the 1960s, computers were purchased to handle credit card applications. In addition, Anderson (2007) mentions that Myers and Forgy proposed using multivariate discriminant analysis for credit rating in 1963. The "US Equal Credit Opportunity Act I" was passed in 1975, bringing full acceptance to credit scoring.

DATA PREPROCESSING:

The data set utilised has 100000 observations and 20 covariates (16 numerical, 7 categorical, and 5 irrelevant columns). It is the German Credit dataset from the UCI machine-learning data collection. Each observation in the dataset corresponds to a single client, with the response showing their actual classification ("Good" or "Bad") and the covariates reflecting different factors relating to their personal or financial information.

The variables in our dataset are listed below along with their variable data type. Our primary target variable from this table will be credit score.

ID	Object
Customer_ID	Object
Month	Object
Name	Object
Age	Object
SSN	Object
Occupation	Object
Annual_Income	Object
Monthly_Inhand_Salary	float64
Num_Bank_Accounts	int64
Num_Credit_Card	int64
Interest_Rate	int64
Num_of_Loan	object
Type_of_Loan	object
Delay_from_due_date	int64
Num_of_Delayed_Payment	object
Changed_Credit_Limit	object
Num_Credit_Inquiries	float64
Credit_Mix	object
Outstanding_Debt	object

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

Credit_Utilization_Ratio	float64
Credit_History_Age	object
Payment_of_Min_Amount	object
Total_EMI_per_month	float64
Amount_invested_monthly	object
Payment_Behaviour	Object
Monthly_Balance	object
Credit_Score	object

In the entire dataset, there are missing values for roughly 8 columns, with Monthly In hand Salary and Type of Loan having the greatest percentages. Missing values have been processed prior to moving on to the initial stage of creating a credit risk assessment model.

There are three ways to deal with missing values:

(1) Delete the samples with incomplete values immediately,

(2) fill in the missing values based on the similarity between the samples.

(3) Fill in the missing values based on the correlation between the variables.

The variable's monthly income loss rate is quite high. As a result, the missing value is filled using the variables' correlation.

Eliminate unnecessary columns like ID, Customer ID, Month, Name, and SSN to tidy up the data. Additionally, alter the datatype of numerical columns like Credit History Age numbers like "22 years and 1 month" include years and months, therefore only years are subtracted from them. Outliers in the dataset should be removed. Our target column, Credit Score, contains a mixture of good, bad, and typical credit classification values.

LITERATURE REVIEW:

Credit Scoring Model based on Improved Tree augmentation Bayesian classification:

The authors of this paper propose a novel feature extraction method for features based on Bayesian classification and enhanced tree augmentation. The features are first reduced to a lower dimension using principal component analysis (PCA), which helps to streamline the inputs to the network. After that, a better Bayesian model is used for classification. The accuracy of the model increased by 2% to 78% after principal component analysis was applied to it. The authors believe that various machine learning algorithms may be used in subsequent work to increase the model's accuracy.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

Credit Scoring Decision Support System:

In this study, the model for building the decision support system was the machine learning algorithm for logistic regression. The suggested decision support system intends to improve the assessment of loan applicants' credit worthiness. In this approach, financial indicators are described as arbitrary characteristics with simulated values. Determine theoretical distributions for the financial indicators, which the decision-maker must do. On this system, the Kolmogorov-Smirnov test was used.

An Empirical Study on Credit Scoring Model for Credit Card by using Data Mining Technology:

In this study, the accuracy of the credit scoring model was examined using 5 different machine learning algorithms. They combined a neural network with a decision tree, logistic regression, a regression tree, and an interaction detector to create the model. Using feature extraction, the primary component that indicates whether or not the consumer has defaulted is first isolated. The accuracy with which the five different models can categorise the dataset is then compared.

Credit scoring model based on BayesianNetwork and Mutual information:

The authors of this research examined feature selection methods like Bayesian Network Mutual Information (BNMI) to reduce the degree of uncertainty among empirical attributes. The learned Bayesian Network was then used to make a robust change in accordance with the shared knowledge. The BNMI model was subsequently put to the test in trials against three various baseline models.

Application of the Hybrid SVM-KNN Modelfor Credit Scoring:

Zhou et al. (2013) employed an ensemble model that combines SVM and the K-Nearest Neighbours algorithm in order to improve the performance of Support Vector Machine in terms of its prediction accuracy. The performance of the combined Support Vector Machine and K-Nearest Neighbours model is superior to that of the separate models. However, it takes a long time to compute the distance function using KNN.

METHODOLOGY:

Logistic regression is a useful classification model that is widely applied to deal with binary classification and multiclassification problems. Logistic regression, which is based on linear regression, often uses the sigmoid function to limit the output to a specific range. In logistic regression, predictions and their probabilities are mapped using a logistic function called the sigmoid function. The sigmoid function, an S-shaped curve, converts any real value into a range between 0 and 1. Additionally, if the output of the sigmoid function (estimated probability) exceeds a predefined threshold on the graph, the model predicts that the instance belongs to that class. If the estimated probability is below the predetermined threshold, the model predicts that the instance does not belong.

For logistic regression, the sigmoid function is known as an activation function and is defined as:

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

$$F(x) = \frac{1}{1 + e^{-x}}$$

where,

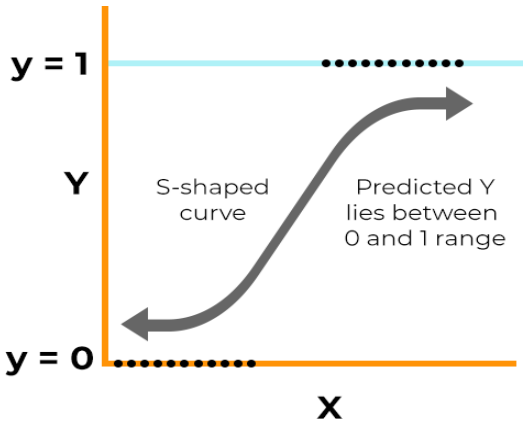
- e = base of natural logarithms
- value = the numerical value that needs to be transformed.

Logistic regression is represented by the following equation:

$$y = \frac{e^{(b_0 + b_1 x)}}{1 + e^{(b_0 + b_1 x)}}$$

where,

- x = input value
- y = predicted output
- b0 = bias or intercept term
- b1 = coefficient for input (x)



Similar to linear regression, this equation linearly combines the input values and utilises weights or coefficient values to predict the output value. In contrast to linear regression, the output value simulated here is a binary value (0 or 1) rather than a numeric value. The logistic regression approach clearly offers advantages when dealing with huge volumes of data, and the gradient descent method considerably speeds up calculation. The model is understandable and reliable after data training, and the expression with actual parameters may be obtained. All the machine learning experiments in this paper are built on the Jupyter Notebook platform and the Python programming language. The Variance Inflation Factor (VIF) in regression analysis determines the degree of multicollinearity. a statistical concept that describes the increase in the variance of a regression coefficient brought on by collinearity. VIF is another well-liked technique for assessing the multicollinearity of a regression model. It establishes the amount by which collinearity raises the variance or standard error of the calculated regression coefficient.

VIF can be calculated by the formula below:

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

$$VIF_i = \frac{1}{1- R_i^2} = \frac{1}{\text{Tolerance}}$$

Ri2 is the uncorrected coefficient of determination for regressing the i^{th} independent variable on the other independent variables in this situation. The inverse of the VIF is referred to as tolerance. Either VIF or tolerance can be used to find multicollinearity, depending on the user's preferences. If Ri2 is zero, it is impossible to estimate the variance of the other independent variables from the i^{th} independent variable. Because there is no correlation between the i^{th} independent variable and the others when VIF or tolerance is 1, this regression model does not exhibit multicollinearity. In this case, the variance of the tenth regression coefficient is not excessive. A VIF above 4 or a tolerance below 0.25 often point to multicollinearity and the need for more study. When the VIF is greater than 10 or the tolerance is less than 0.1, there is a considerable multicollinearity that needs to be corrected.

PROCEDURE OF ANALYSIS:

The steps in logistic regression modelling are asfollows

- Define the problem: To characterise the issue, decide whether it is a binary classification problem by identifying the dependent variable, the independent variables, and the issue.
- Data preparation: Clean up and pre-process the data to make sure it is suitable for logistic regression modelling.
- Exploratory Data Analysis (EDA): Analyse the data for any outliers or anomalies, and depict the connections between the dependent and independent variables.
- Feature selection: Select the independent factors that have a meaningful impact on the dependent variable by eliminating any redundant or irrelevant components.
- Model building: Determine the model's coefficients after training the logistic regression model with the independent variables of choice.
- Model evaluation: To evaluate the performance of the logistic regression model, use the appropriate measures, such as accuracy, precision, recall, F1-score, or AUC-ROC.
- Model improvement: Adjust the model's independent variables, add additional features, or use regularisation techniques in response to the evaluation's findings to reduce overfitting.
- Model deployment: Make predictions using the deployed logistic model based on new data.

RESULTS:

The distributions found in the dataset are as follows:

• The credit utilisation ratio and credit history age are distributed normally.

• Monthly_Inhand_Salary and Delay_from_due_data are right-skewed

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

• Changed_Credit_Limit distribution appears to be in order. We are unable to distribute them effectively because every one of the remaining fields has outliers.

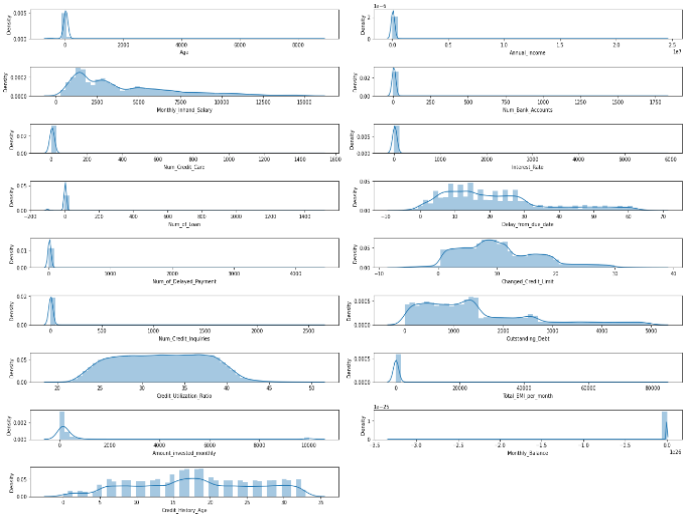


Figure 1: Distribution of numerical columns with outliers

The information in the categories columns is uniformly distributed with the exception of Payment Behaviour and Credit Score. Rows from a variety of job roles are mixed together in the Occupation column. Three categories of information are included in the Credit Mix column. Data from the Yes and No categories are both present in the Payment of Min Amount column. Nevertheless, the category "NM" has to be changed to "No" because it might have been a typo or pre-processing error that was made upstream. Our target score, Credit Score, consists of a mixture of good, bad, and typical credit classification values. More information is available in the "Low spent small values" category of the Payment Behaviour column.

The following columns are connected with the goal variable "Credit score," as shown by the heatmap and histogram. Out of them, there is a strong correlation between Outstanding Debt and Credit Mix.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

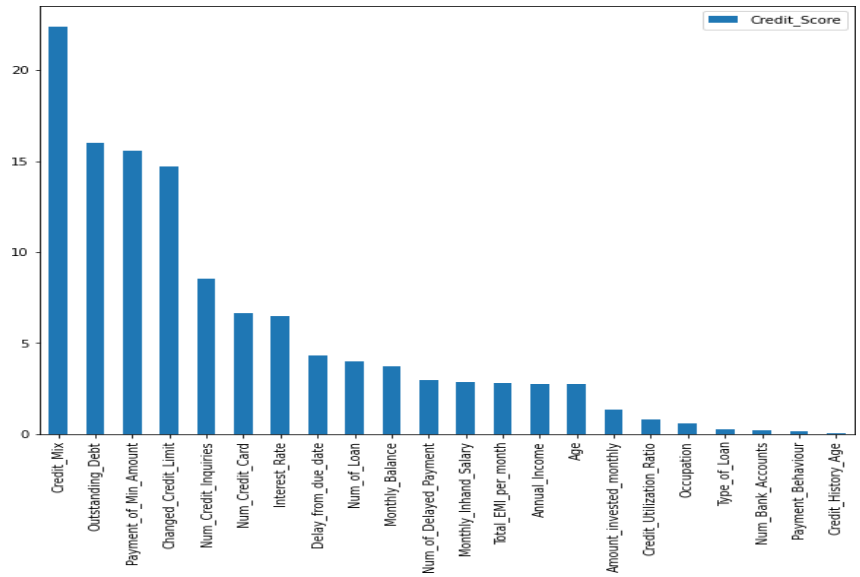


Figure 2: Correlation histogram

The following columns are having the highcorrelation value.

- Credit_Mix - 22.41
- Outstanding_Debt - 16.01
- Payment_of_Min_Amount - 15.55
- Changed_Credit_Limit - 14.70
- Num_Credit_Inquiries - 8.52

The above four columns are the important one because they are highly correlated with the targetvariable.

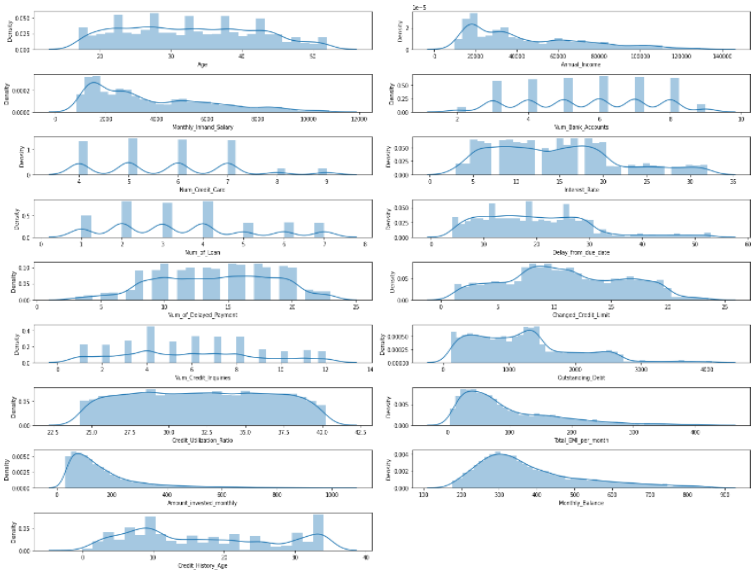


Figure3: Distribution of numeric columns without outliers

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

Final test score values are:

Train Score	Test Score	Outliers in data
0.698	0.686	No
0.178	0.177	Yes

After data analysis, cleaning and training, we were able to get the test score of **68.66** with the help of logistic regression algorithm.

CONCLUSION:

The bulk of banks heavily rely on lending even though it is a risky business for revenue. Banks shouldn't extend credit to clients who can't pay back their debts. Even if banks tighten their lending policies, after a certain amount of time, a certain proportion of credits will inevitably turn into bad loans. By examining the information on non-performing loans (NPLs), it is possible to determine the effectiveness of the credit endorsement procedure. Banks should carefully monitor the advance conceding process and plan a workable credit risk management board. The majority of financial institutions have a centralized department that uses the bank's credit scoring algorithm to assess and categorise financing applications into risky and non-risky customers. This credit scoring process will assess who should obtain credit and how much credit should be issued, with the goal of reducing the risk of loan losses and default rate as a result of the expensive misclassification error. Because of this, the danger grows in direct proportion to the magnitude of the misclassification error.

This essay examines each credit scoring model while using personal credit risk as the research objective. finally integrates the weight of the evidence with the logistic regression model to create a new, accurate prediction model.

A person's data that we obtained includes details such as annual income, credit utilisation ratio, number of accounts, debt and credit history, amount invested, monthly balance, credit score, etc. The target variable, Credit Score, is associated with the following variables. After applying the data cleaning technique to the dataset, we carried out the analysis. In order to determine the correlation between the variables, we used the remove outliers' function to remove the outliers from the data. In order to obtain entirely distinct data, we used a function from the Sklearn module called train test split.

This paper's goal is to give a thorough literature assessment of the theory and use of binary classification algorithms for credit scoring and financial analysis. The overall findings demonstrate the application and importance of the basic credit score calculation techniques as well as some evolution of the logical viewpoint.

REFERENCE:

1. Assef F, Steiner MT, Neto PJS, de Barros Franco DG (2019) Classification

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

algorithms in financial application: credit risk analysis on legal entities. IEEE Lat Am Trans 17(10):1733–1740
2. Ben-David A (1995) Monotonicity maintenance in information-theoretic machine learning algorithms. Mach Learn 19(1):29–43
3. Cornée S (2019) The relevance of soft information for predicting small business credit default: Evidence from a social bank. J Small Bus Manag 57(3):699–719
4. Dastile X, Celik T, Potsane M (2020) Statistical and machine learning models in credit scoring: a systematic literature survey. Appl Soft Comput 106263
5. Davis R, Edelman D, Gammerman A (1992) Machine-learning algorithms for credit-card applications. IMA J Manag Math 4(1):43–51
6. Galindo J, Tamayo P (2000) Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. Comput Econ 15(1/2):107–143
7. Kamalloo E, Saniee Abadeh M (2014) Credit risk prediction using fuzzy immune learning. Adv Fuzzy Syst 2014:1–11
8. Khandani AE, Kim AJ, Lo AW (2010) Consumer credit-risk models via machine-learning algorithms. J Bank Finance 34(11):2767–2787
9. Kozodoi N, Lessmann S, Papakonstantinou K, Gatsoulis Y, Baesens B (2019) A multi-objective approach for profit-driven feature selection in credit scoring. Decis Support Syst 120:106–117
10. Lei K, Xie Y, Zhong S, Dai J, Yang M, Shen Y (2019) Generative adversarial fusion network for class imbalance credit scoring. Neural Comput Appl pp 1–12
11. Li W, Ding S, Chen Y, Wang H, Yang S (2019) Transfer learning-based default prediction model for consumer credit in China. J Supercomput 75(2):862–884
12. Moula FE, Guotai C, Abedin MZ (2017) Credit default prediction modeling: an application of support vector machine. Risk Manag 19(2):158–187
13. Shi J, Sy Zhang, Lm Qiu (2013) Credit scoring by feature-weighted support vector machines. J Zhejiang Univ Sci C 14(3):197–204
14. Siami M, Gholamian MR, Basiri J (2013) An application of locally linear model tree algorithm with combination of feature selection in credit scoring. Int J Syst Sci 45(10):2213–2222
15. Vieira J, Barboza F, Sobreiro VA, Kimura H (2019) Machine learning models for credit analysis improvements: predicting low-income families' default. Appl Soft Comput 83(105):640
16. Wang G, Hao J, Ma J, Jiang H (2011) A comparative assessment of ensemble learning for credit scoring. Expert Syst Appl 38(1):223–230