Credit Scoring Analysis: Case Study of Using Weka

Frane Škegro
Hrvatski Telekom d.d., Croatia
Jovana Zoroja
Faculty of Economics and Business, University of Zagreb, Croatia
Vanja Šimičević
Centre for Croatian Studies, University of Zagreb, Croatia

Abstract

The goal of the paper is to present the overview of methodology of using credit scoring analysis with software Weka. German credit dataset was used in order to develop a decision tree with J.48 algorithm. We present characteristics of the dataset and the main results with the focus to the interpretation of Weka output. Paper could be useful for the users of Weka that aim to use it for credit scoring analysis.

Keywords: data base, credit risk, data mining, knowledge discovery, granting credits
JEL classification: C80, D81

Introduction

The paper presents one usage of Weka software for credit scoring, using data mining approach to uncover hidden trends and to make accuracy based predictions. In order to fulfil the goal, we use hypothetical German credit data set available on UCI Machine Learning Repository which contains sample of 1000 debtors classified as „good“ or „bad“ (UCI Machine Learning Repository, German data set). One of the most popular data mining techniques, decision tree algorithm J.48, is applied to build prediction models.

This paper consists of four sections including Introduction part as the first one. The second section presents the research methodology including data description and methodology used. The third section provides given results. Finally, the last section concludes the paper, including the limitations of the study and future implications.

Methodology

In the second section of the paper we describe data which we have used and how we have analysed it. Therefore, we present data set regarding German credit data and decision tree, one of the data mining methods which are usually used for classification problems.

Data

In this study we have used German credit data set. There are information about 1000 debtors which are described with 20 variables (7 numeric and 13 nominal) and which are classify as “good” or “bad”. Data set is available at UCI Machine Learning Repository. We have analysed five attributes of model regarding German data seta, and we have explained each attribute with the last one, class attribute (Table 1).

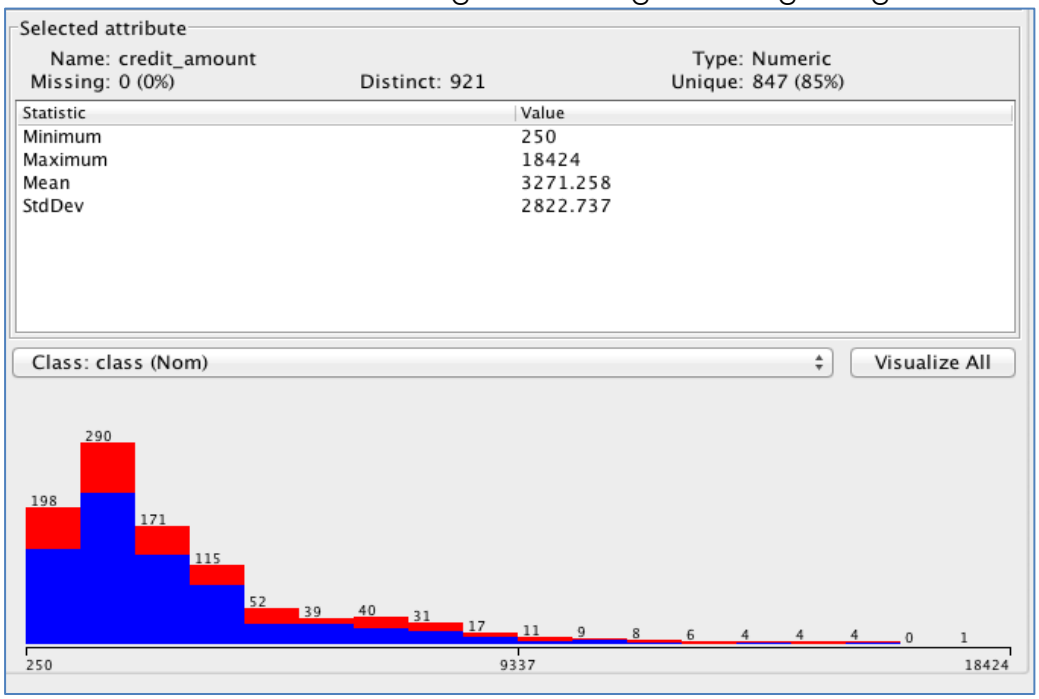
Two variables will be analysed for the demonstration purposes (Figure 1 and Figure 2). Minimal amount of approved credit was 250,00 DM and maximum amount was 18.424,00 DM. Results showed that the highest number of bank clients borrowed smaller amount of credit. Approximately 20% of clients (198) borrowed between 250,00 and 1.259,00 DM, and only 25% were classify as „bad" debtors. The highest amount of credit borrowed the smaller number of clients. Only four bank clients borrowed between 15.395,00 and 16.404,00 DM and three of them are classify as „good" debtors.

Table 1
List of variables

Name of Variable	Type of Variable	Description
Checking status	Nominal	A11: ... < 0 DM; A12: 0 < ... < 200 DM A13: ... >= ... 200 DM; A14: no account
Credit amount	Numeric	Min=250; Max=18424; Mean=3271.258; St.Dev.=2822.737
Employment	Nominal	A71: unemployed; A72: ... < 1 year A73: ... <= ... 4 year; A74: 4 <= ... 7 year A75: ... >= 7 year
Existing credit	Numeric	Min=1; Max=4; Mean=1.407; St.Dev.=0.578
Class	Nominal	1="good"=700; 2="bad"=300

Source: Authors' survey

Figure 1
Evaluation of clients' credit rating for returning credit regarding credit amount

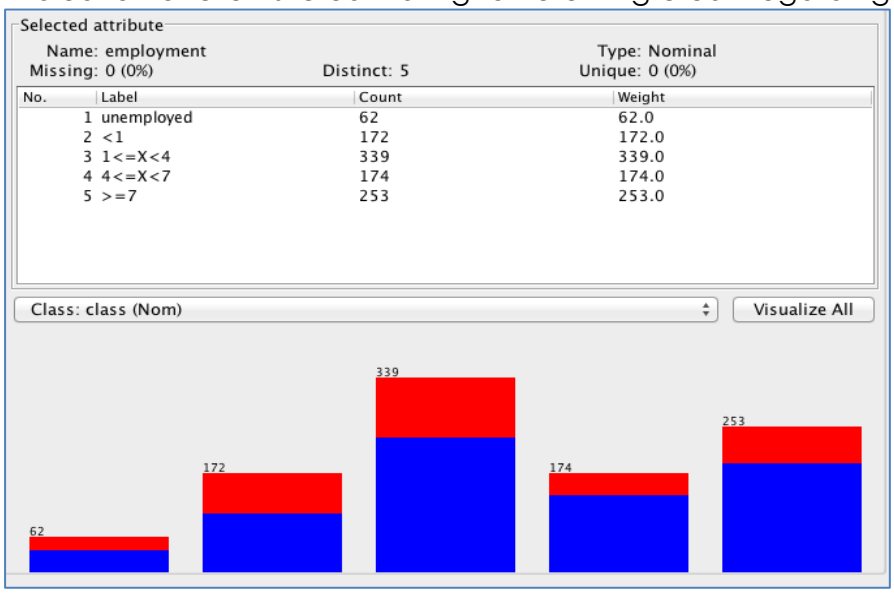


Source: Authors' survey, Weka

Bank clients are classified into five categories regarding employment: unemployed, employment less than one year, employees between one and four years, employees between four and seven years and employees more than seven years.

Results showed that clients with longer employ period are classifying as good „debtors"(Figure 2). The lowest numbers of clients are unemployed (62) and most of them are classifying as „good" debtors. The highest number of clients works more than seven years (253), and more than 60% of them are classify as „good" debtors.

Figure 2
Evaluation of clients' credit rating for returning credit regarding employment



Source: Authors' work, Weka

Decision tree

Decision trees present one of the several classification methods. The main goal of the decision trees method is to group variables into one or more categories according to the target attributed (Yap et al., 2011; Chuang, Chia, Wong, 2013). Following prerequisites are needed to use decision trees method: (i) previously defined final number of categories for each variable, (ii) each data should be part of only one category, (iii) large data set in order to have at least ten observations for each group. Decision trees method offers many different algorithms. Therefore, it is important to select appropriate algorithm for analysis.

There are several benefits of decision trees method: simple usage and implementation, efficient and objective analysis, easy to understand and interpret obtained results, usage of qualitative and quantitative data (Patel, Sarvakar, 2014; Olson, Chae, 2012). However, one of possible obstacle is high variance in decision tree analysis.

Results

Table 2 presents compendious classification model of decision tree analysis where J48 algorithm has been used. Obtained results are shown textually (Table 2) and graphically (Figure 3).

Following attributes are used in decision tree analysis: Checking status, Credit amount, Employment, Existing credits and Class. Obtained results have shown that checking status is class attribute, size of tree is 20 and number of leaves is 13.

Each attribute have different values. For example, attribute Checking status have four possible values: '< 0' (negative current account balance), '0<=x=<200' (current

account balance up to 200 DM), '>=200' (current account balance higher than 200 DM) 'no checking status' (clients who do not have current account).

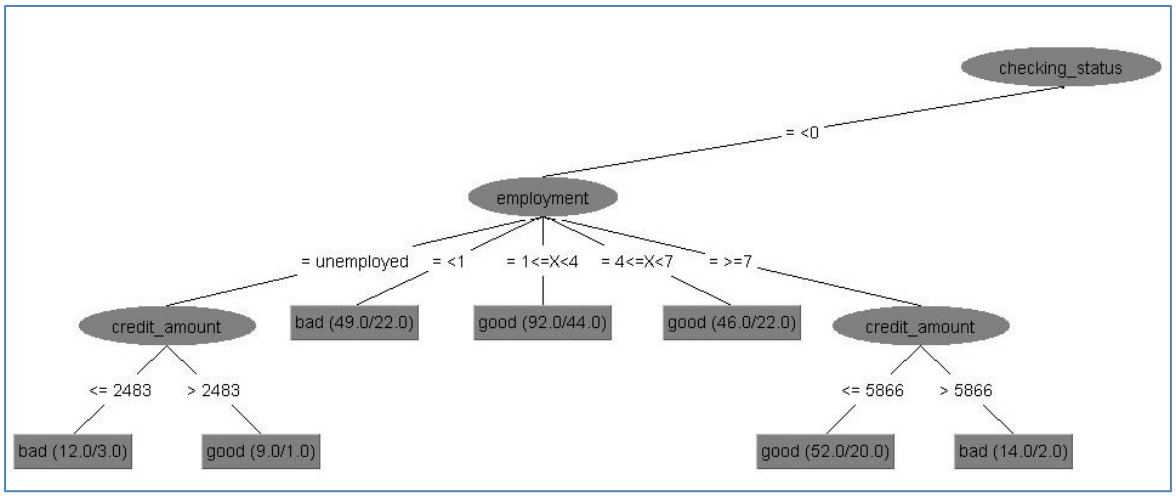
Table 2
Decision tree analysis for good and bad debtors

```
J48 pruned tree
------------------

checking_status = <0
|   employment = unemployed
|   |   credit_amount <= 2483: bad (12.0/3.0)
|   |   credit_amount > 2483: good (9.0/1.0)
|   employment = <1: bad (49.0/22.0)
|   employment = 1<=X<4: good (92.0/44.0)
|   employment = 4<=X<7: good (46.0/22.0)
|   employment = >=7
|   |   credit_amount <= 5866: good (52.0/20.0)
|   |   credit_amount > 5866: bad (14.0/2.0)
checking_status = 0<=X<200
|   credit_amount <= 9857: good (249.0/88.0)
|   credit_amount > 9857
|   |   existing_credits <= 1
|   |   |   credit_amount <= 12204: good (4.0/1.0)
|   |   |   credit_amount > 12204: bad (11.0)
|   |   existing_credits > 1: bad (5.0)
checking_status = >=200: good (63.0/14.0)
checking_status = no checking: good (394.0/46.0)

Number of Leaves  :     13

Size of the tree :      20


Time taken to build model: 0.08 seconds
```

Source: Authors' work, Weka

Figure 3 presents one part of the whole decision tree where attribute Checking status branched on attribute Employment, regarding those clients who have negative current account balance, with five different possibilities: 'Unemployed', '=<1' (employed less than a year), '1<=X<4' (employed exactly one year or more than one year but until four years), '4<=<7' (employed exactly four years or more than four years but until seven years), '>=7' (employed more than seven years).

Figure 3
Example of allocation of class value "Good" or "Bad" for attribute Employment



Source: Authors' work, Weka

After second, third and fourth modality there is decision about client (good or bad debtor) without data about credit amount. For example, "good (92.0/44.0)" mean that there are correctly classified 92 clients, who have negative current account balance and are employed, as good debtors. Other 44 clients are good debtors who are inaccurately predicted as "bad" debtors.

First and fifth value for attribute Employment leads to attribute Credit amount. For example, "good (52.0/20.0)" mean that there are correctly classified 52 clients, who have negative current account balance, are employed and asked for credit amount more than 5866 DM, as good debtors. Other 20 clients are bad debtors who are inaccurately predicted as "good" debtors.

Table 3 presents classification efficiency measures. Total number of instances is 1000 and in this presented model 70% of instances is correctly classified.

Table 3
Classification efficiency measures

```
Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         700               70      %
Incorrectly Classified Instances       300               30      %
Kappa statistic                          0.1339
Mean absolute error                      0.3603
Root mean squared error                  0.4351
Relative absolute error                 85.7465 %
Root relative squared error             94.9432 %
Total Number of Instances             1000
```

Source: Authors' work, Weka

Table 4 presents confusion matrix for decision tree analysis. In models with class attribute and two modalities "good" and "bad", one prediction could have four possible results: *True positive (TP), True negative (TN), False negative (FN) and False positive (FP)*.

Table 4
Confusion Matrix for decision tree analysis

```
=== Confusion Matrix ===

   a   b   <-- classified as
 642  58 |   a = good
 242  58 |   b = bad
```

Source: Authors' work, Weka

Results of the"Confusion Matrix" showed that 700 instances were correctly classified (642+58) and 300 (the rest, out of 1000) were incorrectly classified. More precisely, *True positive* has 642 clients who are really good debtors. *True negative* has 58 clients who are classified as "bad" debtors (300 clients are classified as "bad" clients). False negative and False positive are wrong classifications. *False negative* presents 58 clients who are inaccurately predicted as "bad" debtors, while *False positive presents* 242 clients who are inaccurately predicted as "good" debtors.

Conclusion

According to the previous research we can conclude that there have been many positive changes in banking sector in the 21st century. Banking sector follow the new trends and development in information technology using advanced techniques, such as data mining. In this paper the usage of Weka software for credit scoring was presented on the case study of German credit dataset. The hypothetical results indicate that there is higher number of clients classified as „good", clients who are paying their credit on time. Results also indicate that those clients who have higher amount on their current account and who are working for longer period of time are better debtor.

References

1. Chuang, Y.F., Chia, S.H., Wong, J.Y. (2013), "Customer Value Assessment of Pharmaceutical Marketing in Taiwan", Industrial Management and Data Systems, Vol. 113 No. 9, pp. 1315-1333.
2. Olson, D.L., Chae, B. (2012), "Direct Marketing Decision Support through Predictive Customer Response Modelling", Decision Support Systems, Vol. 54 No. 1, pp. 443–451.
3. Patel, H.G., Sarvakar, K. (2014), "Research Challenges and Comparative Study of Various Classification Technique Using Data Mining", International Journal of Latest Technology in Engineering, Management & Applied Science, Vol. 3 No. 9, pp. 170-176.
4. UCI Machine Learning Repository, German data set, available at: https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data) (21 March 2017)
5. Yap, B.W., Ong, S.H., Husain, N.H.M. (2011), "Using Data Mining to Improve Assessment of Credit Worthiness via Credit Scoring Models", Expert Systems with Applications, Vol. 38 No. 10, pp. 13274-1328.

About the authors

Frane Škegro, mag.oec. is currently Marketing proposition manager in Hrvatski Telekom d.d. Author has been working in HT since 2013, when he graduated on Graduate study Managerial Informatics at Faculty of Economics & Business, University of Zagreb. His research interests are IT related innovation, big data, video and gaming industry. Author can be contacted at **franeskegro@gmail.com.**

Jovana Zoroja, Ph.D. is an Assistant Professor at the Faculty of Economics and Business, University of Zagreb, Department of Informatics where she received PhD in Information Systems. She was also educated at the LSE – Summer School in London in the field of Business Development and ICT Innovation. Her main research interests are information and communication technology, e-learning, simulation games and simulation modelling. She is actively engaged in number of projects (FP7-ICT, bilateral cooperation, national projects, Erasmus). Jovana Zoroja published several scientific papers in international and national journals and participated in many scientific international conferences. Author can be contacted at **jzoroja@efzg.hr.**

Vanja Šimičević has PhD in Economics from the University of Zagreb, Faculty of Economics and Business in the area of quantitative economics. Her major area of research is focused on applications of quantitative methods in social sciences and on those topics she published number of papers. She is Associate Professor at the University of Zagreb Centre for Croatian Studies, Head of Sociology Department, teaching Multivariate Statistical Methods, and Statistics in Social Sciences. Author can be contacted at **vanja.simicevic@zg.htnet.hr.**