

Determining customer limits by data mining methods in credit allocation process

Tuğçe Ayhan¹, Tamer Uçar²

¹Information Technologies, Graduate School, Bahçeşehir University, İstanbul, Turkey

²Department of Software Engineering, Faculty of Engineering and Natural Sciences, Bahçeşehir University, İstanbul, Turkey

Article Info

Article history:

Received Dec 30, 2020

Revised Oct 8, 2021

Accepted Oct 25, 2021

Keywords:

Banking

Credit allocation process

Data mining

Machine learning algorithms

ABSTRACT

The demand for credit is increasing constantly. Banks are looking for various methods of credit evaluation that provide the most accurate results in a shorter period in order to minimize their rising risks. This study focuses on various methods that enable the banks to increase their asset quality without market loss regarding the credit allocation process. These methods enable the automatic evaluation of loan applications in line with the sector practices, and enable determination of credit policies/strategies based on actual needs. Within the scope of this study, the relationship between the predetermined attributes and the credit limit outputs are analyzed by using a sample data set of consumer loans. Random forest (RF), sequential minimal optimization (SMO), PART, decision table (DT), J48, multilayer perceptron (MP), JRip, naïve Bayes (NB), one rule (OneR) and zero rule (ZeroR) algorithms were used in this process. As a result of this analysis, SMO, PART and random forest algorithms are the top three approaches for determining customer credit limits.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Tamer Uçar

Department of Software Engineering, Faculty of Engineering and Natural Sciences, Bahçeşehir University

Beşiktaş, İstanbul, 34349, Turkey

Email: tamer.ucar@eng.bau.edu.tr

1. INTRODUCTION

Data analysis and data mining techniques are used in many domains to convert raw data into knowledge [1]. These methods are applied on various fields such as medical diagnosis [2], travel recommenders [3]–[5], fraud detectors [6], numerous classification problems [7] and many more. Banking domain is one of these fields with a wide range of data analysis needs [8]. Credit assessment process is an important data mining application area in this domain.

One of the main activities of a bank is credit lending. In order to improve market share and increase sales profitability, banks need to develop a good credit assessment process in which they can carry out lending [9]. This is an important requirement for banks due to the constantly evolving market and increasing competition. The credit assessment process mainly aims to make an analysis to determine whether the party requesting the loan has the power to fulfill its obligation to repay the loan at the end of the loan agreement and to reduce the likelihood of non-payment of the loan as much as possible by determining whether it has the willingness to pay the loan [10].

Developing a unique credit assessment system which produces accurate, stable and reliable results in a short time is an important advantage in competitive market conditions. Additionally, such a unique credit assessment system, which should be equipped with safe techniques, should help the bank achieve its strategic goals in the corporate field and create user satisfaction by meeting the expectations of the customers. For this

reason, many banks use data mining methods to extract meaningful information from customer data and try to manage the loan evaluation process by taking this information as a core reference [11]–[17].

In this study, the characteristics of people requesting credit were evaluated and the effects of these features on the available credit limit was examined. The main purpose of this study is to propose a decision-making approach to minimize the risk of non-repayment of loans. In order to achieve this goal, a sample data set containing attributes related to credit allocation is used. The relationship between the data set and the credit limit, which was harmonized with data mining methods, was examined with random forest (RF), sequential minimal optimization (SMO), PART, decision table (DT), J48, multilayer perceptron (MP), JRip, naïve Bayes (NB), one rule (OneR) and zero rule (ZeroR) algorithms. The obtained metrics of the produced prediction models were compared and analyzed. The models with the most accurate prediction results were proposed as possible decision-making approaches in credit allocation process. The rest of the paper is organized as follows: section 2 describes the data gathering process and includes a background of the data mining algorithms used in this research. Section 3 presents the obtained results and section 4 contains conclusion and future plans of this study.

2. RESEARCH METHOD

A total of 10 different data mining algorithms were used in this research to find the best approach for determining customer credit limits. The data set and applied data mining algorithms are described in detail in the following subsections.

2.1. Data gathering and processing

The data set reflecting the characteristics of possible customer credit requests and credit limits was generated by a banking specialist. 401 records were collected to run data experiments. It has 14 input attributes derived from 4 main attribute groups and it has one output attribute which is the corresponding credit limit. Table 1 lists attribute groups, attributes and their descriptions. Credit limit attribute was discretized into 7 categories based on its data range. Table 2 lists credit limit categories by minimum and maximum credit values.

Table 1. Data set attributes

Attribute groups	Definition	Attributes
Monthly income-installment ratio (MIIR)	Customer's ratio of monthly income over installment.	- Up to 50% - Between 50% and 60% - Between 60% and 70% - Between 70% and 80%
Income type	Customer's income type.	- Private sector paid worker - Private sector unpaid worker - Public sector paid worker - Public sector unpaid worker
Investigation result level	Customer's status about previous credit payments.	- Reject (Bad reputation according to past payment history) - Unclear - Accept (Good reputation according to past payment history)
Risk level	Customer's risk level based on investigation result level and previous banking history.	- High - Medium - Low
Credit limit	Customer's credit limit based on Monthly Income-Installment Ratio, Income Type, Investigation Result Level and Risk Level attributes.	- Credit Limit

Table 2. Credit limit categories

Credit limit category	Minimum credit value (in TL)	Maximum credit value (in TL)
Group 1	0	8200
Group 2	8201	16400
Group 3	16401	24600
Group 4	24601	32800
Group 5	32801	41000
Group 6	41001	49200
Group 7	49201	82000

Preprocessed data set was converted to attribute-relation file format (ARFF) to run data mining algorithms. Each numeric attribute in the ARFF data file may contain 0 or 1 where 1 represents true and 0 represents false. Figure 1 shows the structure of the ARFF data file. Info gain attribute eval algorithm [18] is used to measure the information gain of each attribute against the credit limit attribute (output class). According to the attribute rankings unclear investigation result level attribute has no effect on decreasing the overall entropy, therefore it has been removed from the data set. Table 3 lists information gain attribute rankings.

```
@relation data

@attribute MIIR_50 numeric
@attribute MIIR_5060 numeric
@attribute MIIR_6070 numeric
@attribute MIIR_7080 numeric
@attribute PublicSector_Paid numeric
@attribute PublicSector_Unpaid numeric
@attribute PrivateSector_Paid numeric
@attribute PrivateSector_Unpaid numeric
@attribute LowRiskLevel numeric
@attribute MediumRiskLevel numeric
@attribute HighRiskLevel numeric
@attribute InvestigationAccept numeric
@attribute InvestigationUnclear numeric
@attribute InvestigationReject numeric
@attribute CreditLimit {CAT1,CAT2,CAT3,CAT4,CAT5,CAT6,CAT7}
```

Figure 1. Data file structure

Table 3. Information gain attribute rankings

Attribute Name	Attribute Rank
MIIR_7080	0.608
MIIR_50	0.475
PrivateSector_Unpaid	0.449
MIIR_6070	0.438
InvestigationAccept	0.4
InvestigationReject	0.367
LowRiskLevel	0.283
HighRiskLevel	0.226
PrivateSector_Paid	0.213
PublicSector_Paid	0.197
MIIR_5060	0.183
MediumRiskLevel	0.154
PublicSector_Unpaid	0.13
InvestigationUnclear	0

2.2. Data mining algorithms

RF, SMO, PART, DT, J48, MP, JRip, NB, OneR and ZeroR algorithms were used for building prediction models for classification of customer credit limits. Brief descriptions of these algorithms are listed:

- Random forest (RF): RF is an ensemble learning method which is a mixture of several tree-based predictors. Output of the algorithm is the mode of the classification classes that the forest contains [19].
- Support vector machines (SVM) and sequential minimal optimization (SMO): SVM is a supervised learning approach for classification tasks in machine learning domain. The algorithm tries to find a hyperplane to classify data points in a given data set distinctly. SMO is an optimization in SVM which solves the quadratic programming problem in SVM model training phase [20], [21].
- PART: Generates classification rules by creating partial decision trees without global optimization. The algorithm is an efficient separate-and-conquer rule learning technique [22].
- Decision table (DT): DT is a simple but effective supervised learning algorithm used in classification problems. It has a collection of rules called decision list which is used in the classification process. Each rule is processed sequentially until a matching rule is found [23].
- J48: J48 is a similar version of C4.5 decision tree algorithm which is implemented using Java programming language. The algorithm is capable of generating a pruned/unpruned decision tree based on information entropy [24].

- Multilayer perceptron (MP): MP is a feed-forward artificial neural network which uses backpropagation for training classification models [25].
- JRip: It is the implementation of repeated incremental pruning to produce error reduction (RIPPER) approach which is a rule-based classifier [26].
- Naïve bayes (NB): NB is probabilistic classification approach based on Bayes' theorem. It assumes that a feature of a class is independent of any other feature [27].
- One rule (OneR): It is a simple classification algorithm which tries to find the one rule with the minimum prediction error according to a training data set [28].
- Zero rule (ZeroR): This algorithm is another simple classification approach which ignores predictors other than the target (class) attribute. The mean is calculated for numeric classes and mode is computed for nominal classes [29]. This study uses ZeroR for determining the baseline performance for machine learning algorithms.

2.3. Comparing algorithms

Accuracy, precision and recall metrics are computed for comparing data mining models used in this study. These metrics are calculated using true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values extracted from confusion matrices of prediction models. A confusion matrix holds the number of actual values in its rows and keeps the number of predicted values in its columns.

TP is the number of values where the classification model correctly predicts the positive class instances. In a similar way, TN is the number of values where the classification model correctly predicts the negative class instances. FP represents the number of incorrectly predicted positive class instances whereas, FN is the number of values where the classification model incorrectly predicts the negative class instances. Accuracy, precision and recall scores are calculated according to the described TP, TN, FP and FN values. Accuracy is the percentage of correctly predicted instances. Precision is the ratio of TP values divided by the sum of TP and FP values. Recall is the ratio of TP values divided by the sum of TP and FN values.

2.4. Building classification models

Waikato environment for knowledge analysis (WEKA) is an open-source machine learning platform [30]. It is used for training and testing the classification models with the described bank-customer data set. Each model is tested with both 10-fold cross-validation and 66% percentage-split methods. Comparison metrics are collected for both of these testing approaches and results are discussed in the next section.

3. RESULTS AND DISCUSSION

The algorithms mentioned in section 2.2 are applied on the final version of the banking data set. Two test experiments are conducted based on two different testing approaches. Accuracy, precision and recall metrics are calculated based on these experiments. Obtained results are listed in Table 4.

Table 4. Performance results of classification models

Algorithm	Experiment 1: 10-Fold Cross-Validation			Experiment 2: 66% Split Test		
	Accuracy (%)	Precision	Recall	Accuracy (%)	Precision	Recall
RF	96.76	0.97	0.97	93.38	0.94	0.93
SMO	96.51	0.97	0.97	93.38	0.95	0.93
PART	96.51	0.97	0.97	93.38	0.95	0.93
DT	95.51	0.96	0.96	90.44	0.92	0.90
J48	95.51	0.96	0.96	91.91	0.94	0.92
MP	95.26	0.95	0.95	91.91	0.94	0.92
JRip	94.51	0.95	0.95	91.91	0.94	0.92
NB	86.03	0.90	0.86	84.56	0.88	0.85
OneR	44.14	0.55	0.44	38.24	0.49	0.38
ZeroR	26.69	0.26	0.26	22.06	0.22	0.22

According to the performance results of classification models, RF algorithm has the best classification accuracy and recall scores in both of the experiments. And it has the best classification precision score in Experiment 1 whereas SMO and PART algorithms have the highest classification precision score in Experiment 2. All classification models produced better scores than ZeroR baseline performance.

Table 5 shows comparison results of prediction models. Each row in the table starts with a prediction model. The columns after the prediction model list the names of other algorithms that produced lower results than the model in that row according to the accuracy, precision and recall scores. These results are listed for both experiments. The last column contains information about how many times the prediction

model in the row produces better results than other algorithms. For example, SMO and PART has better results in 43 comparisons with other algorithms based on the results of both experiments whereas RF has better results in 41 comparisons. J48, DT, MP, JRip, NB, OneR and ZeroR algorithms have no better comparison results than 27. Based on the presented results in Tables 4 and 5, RF algorithm has the highest scores in most of the metrics for both of the experiments whereas SMO and PART algorithms have the greatest number of better comparison results.

Table 5. Comparison results of classification models

Algorithm	Experiment 1: 10-Fold Cross-Validation			Experiment 2: 66% Split			Number of Better Results
	Has Better Accuracy Than	Has Better Precision Than	Has Better Recall Than	Has Better Accuracy Than	Has Better Precision Than	Has Better Recall Than	
SMO	DT, J48, MP, JRip, NB, OneR, ZeroR	DT, J48, MP, JRip, NB, OneR, ZeroR	DT, J48, MP, JRip, NB, OneR, ZeroR	J48, MP, JRip, DT, NB, OneR, ZeroR	RF, J48, MP, JRip, DT, NB, OneR, ZeroR	J48, MP, JRip, DT, NB, OneR, ZeroR	43
PART	DT, J48, MP, JRip, NB, OneR, ZeroR	DT, J48, MP, JRip, NB, OneR, ZeroR	DT, J48, MP, JRip, NB, OneR, ZeroR	J48, MP, JRip, DT, NB, OneR, ZeroR	RF, J48, MP, JRip, DT, NB, OneR, ZeroR	J48, MP, JRip, DT, NB, OneR, ZeroR	43
Random Forest	SMO, PART, DT, J48, MP, JRip, NB, OneR, ZeroR	DT, J48, MP, JRip, NB, OneR, ZeroR	DT, J48, MP, JRip, NB, OneR, ZeroR	J48, MP, JRip, DT, NB, OneR, ZeroR	DT, NB, OneR, ZeroR	J48, MP, JRip, DT, NB, OneR, ZeroR	41
J48	MP, JRip, NB, OneR, ZeroR	MP, JRip, NB, OneR, ZeroR	MP, JRip, NB, OneR, ZeroR	DT, NB, OneR, ZeroR	DT, NB, OneR, ZeroR	DT, NB, OneR, ZeroR	27
Decision Table	MP, JRip, NB, OneR, ZeroR	MP, JRip, NB, OneR, ZeroR	MP, JRip, NB, OneR, ZeroR	NB, OneR, ZeroR	NB, OneR, ZeroR	NB, OneR, ZeroR	24
Multilayer Perceptron	JRip, NB, OneR, ZeroR	NB, OneR, ZeroR	NB, OneR, ZeroR	DT, NB, OneR, ZeroR	DT, NB, OneR, ZeroR	DT, NB, OneR, ZeroR	22
JRip	NB, OneR, ZeroR	NB, OneR, ZeroR	NB, OneR, ZeroR	DT, NB, OneR, ZeroR	DT, NB, OneR, ZeroR	DT, NB, OneR, ZeroR	21
Naive Bayes	OneR, ZeroR	OneR, ZeroR	OneR, ZeroR	OneR, ZeroR	OneR, ZeroR	OneR, ZeroR	12
OneR	ZeroR	ZeroR	ZeroR	ZeroR	ZeroR	ZeroR	6
ZeroR	None	None	None	None	None	None	0

4. CONCLUSION

This study shows a detailed data mining algorithm comparison for customer credit allocation process using two different experiment sets. RF, SMO, PART, DT, J48, MP, JRip, NB, OneR and ZeroR algorithms are trained and tested on a banking data set which has characteristics of possible customer credit requests and credit limits. Obtained test results suggest that SMO, PART and RF algorithms are the most accurate three data mining approaches for customer credit allocation process. Proposing a hybrid data mining solution using SMO, PART and RF algorithms for the implementation of a customer credit allocation tool is planned as a future study.




REFERENCES

- [1] J. Han, J. Pei, and M. Kamber, *Data mining*. Elsevier, 2012.
- [2] T. Uçar, A. Karahoca, and D. Karahoca, "Tuberculosis disease diagnosis by using adaptive neuro fuzzy inference system and rough sets," *Neural Computing and Applications*, vol. 23, no. 2, pp. 471–483, Aug. 2013, doi: 10.1007/s00521-012-0942-1.
- [3] T. Uçar, A. Karahoca, and D. Karahoca, "NTRS: A new travel recommendation system framework by hybrid data mining," *International Journal of Mechanical Engineering and Technology (IJMET)*, vol. 10, no. 1, pp. 935–946, 2019.
- [4] I. Y. Choi, Y. U. Ryu, and J. K. Kim, "A recommender system based on personal constraints for smart tourism city," *Asia Pacific Journal of Tourism Research*, vol. 26, no. 4, pp. 440–453, Apr. 2021, doi: 10.1080/10941665.2019.1592765.
- [5] L. Santamaria-Granados, J. F. Mendoza-Moreno, and G. Ramirez-Gonzalez, "Tourist recommender systems based on emotion recognition-a scientometric review," *Future Internet*, vol. 13, no. 1, Dec. 2020, Art. no. 2, doi: 10.3390/fi13010002.
- [6] J. Perols, "Financial statement fraud detection: an analysis of statistical and machine learning algorithms," *AUDITING: A Journal of Practice and Theory*, vol. 30, no. 2, pp. 19–50, May 2011, doi: 10.2308/ajpt-50009.
- [7] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification," *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–40, Jun. 2021, doi: 10.1145/3439726.
- [8] H. Bekamiri, S. F. Ghasempour Ganji, B. Simonetti, and S. A. H. Seno, "A new model to identify the reliability and trust of internet banking users using fuzzy theory and data-mining," *Mathematics*, vol. 9, no. 9, Apr. 2021, Art. no. 916, doi: 10.3390/math9090916.




- [9] S.-B. Tsai, G. Li, C.-H. Wu, Y. Zheng, and J. Wang, "An empirical research on evaluating banks' credit assessment of corporate customers," *SpringerPlus*, vol. 5, no. 1, pp. 1-13, Dec. 2016, Art. no. 2088, doi: 10.1186/s40064-016-3774-0.
- [10] J. Witzany, "Credit risk management," in *Credit Risk Management*, Cham: Springer International Publishing, 2017, pp. 5–18.
- [11] I. G. N. N. Mandala, C. B. Nawangpalupi, and F. R. Praktiko, "Assessing credit risk: an application of data mining in a rural bank," *Procedia Economics and Finance*, vol. 4, pp. 406–412, 2012, doi: 10.1016/S2212-5671(12)00355-3.
- [12] B. W. Yap, S. H. Ong, and N. H. M. Husain, "Using data mining to improve assessment of credit worthiness via credit scoring models," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13274–13283, Sep. 2011, doi: 10.1016/j.eswa.2011.04.147.
- [13] M. de M. Sousa and R. S. Figueiredo, "Credit analysis using data mining: application in the case of a credit union," *Journal of Information Systems and Technology Management*, vol. 11, no. 2, pp. 379–396, Aug. 2014, doi: 10.4301/S1807-17752014000200009.
- [14] S. Khemakhem and Y. Boujelbene, "Predicting credit risk on the basis of financial and non-financial variables and data mining," *Review of Accounting and Finance*, vol. 17, no. 3, pp. 316–340, Aug. 2018, doi: 10.1108/RAF-07-2017-0143.
- [15] A. Hooman, G. Marthandan, W. F. W. Yusoff, M. Omid, and S. Karamizadeh, "Statistical and data mining methods in credit scoring," *The Journal of Developing Areas*, vol. 50, no. 5, pp. 371–381, 2016, doi: 10.1353/jda.2016.0057.
- [16] F. N. Koutanaei, H. Sajedi, and M. Khanbabaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring," *Journal of Retailing and Consumer Services*, vol. 27, pp. 11–23, Nov. 2015, doi: 10.1016/j.jretconser.2015.07.003.
- [17] M. Abedini, F. Ahmadzadeh, and R. Noorossana, "Customer credit scoring using a hybrid data mining approach," *Kybernetes*, vol. 45, no. 10, pp. 1576–1588, Nov. 2016, doi: 10.1108/K-09-2015-0228.
- [18] J. Novakovic, "Using information gain attribute evaluation to classify sonar targets," in *17th Telecommunications forum TELFOR 2009*, 2009, pp. 1351–1354.
- [19] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [20] K.-R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," in *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 1997, pp. 999–1004.
- [21] J. C. Platt, "Sequential minimal optimization: a fast algorithm for training support vector machines," *Support Vector Machines*, 1998.
- [22] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," 1998.
- [23] R. Kohavi, "The power of decision tables," in *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 1995, pp. 174–189.
- [24] S. L. Salzberg, "C4.5: programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," *Machine Learning*, vol. 16, no. 3, pp. 235–240, Sep. 1994, doi: 10.1007/BF00993309.
- [25] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, and classification," *IEEE Transactions on Neural Networks*, vol. 3, no. 5, pp. 683–697, 1992, doi: 10.1109/72.159058.
- [26] W. W. Cohen, "Fast effective rule induction," in *Machine Learning Proceedings 1995*, Elsevier, 1995, pp. 115–123.
- [27] I. Rish, "An empirical study of the naive Bayes classifier," *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, pp. 41–46, 2001.
- [28] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, pp. 63–90, 1993.
- [29] I. H. Witten, E. Frank, and M. A. Hall, *Data mining: practical machine learning tools and techniques*. Elsevier, 2011.
- [30] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, Nov. 2009, doi: 10.1145/1656274.1656278.

BIOGRAPHIES OF AUTHORS



Tuğçe Ayhan    holds a master's degree in Information Technologies from Bahçeşehir University. She works as a Business Analyst in Commercial Credits/Loans team in banking sector. E-mail: tgcaihan89@gmail.com.



Dr. Tamer Uçar    holds a PhD in Computer Engineering. His research interests are software development, data mining, recommender systems, medical data analysis and big data. He has published articles about data mining applications in health and tourism domains. He is currently working for Bahçeşehir University Software Engineering Department as full-time faculty. E-mail: tamer.ucar@eng.bau.edu.tr.