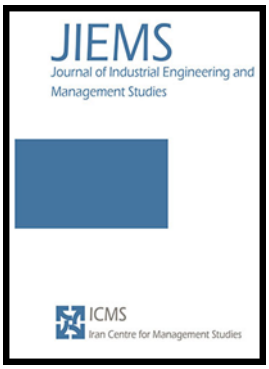
JIEMS
Journal of Industrial Engineering and Management Studies

Vol. 5, No. 2, 2018, pp. 1-12

DOI: 10.22116/JIEMS.2018.80681

www.jiems.icms.ac.ir



Investigating the missing data effect on credit scoring rule based models: The case of an Iranian bank

Seyed Mahdi Sadatrasoul[1,*], Zeynab Hajimohammadi[2]

Abstract
Credit risk management is a process in which banks estimate probability of default (PD) for each loan applicant. Data sets of previous loan applicants are built by gathering their data, and these internal data sets are usually completed using external credit bureau's data and finally used for estimating PD in banks. There is also a continuous interest for bank to use rule based classifiers to build their default prediction models. However, in practice the data records are usually incomplete and have some missing values and this make problems for banks, especially in credit risk portfolios which are low default and makes model rule based building complex. Several strategies could be used in order to handle the missing data issue. This paper used five missing value handling strategies including; ignoring, replacing with random, mean, C&R tree induced values and elimination strategies in a real credit scoring dataset. Experimental results show that ignoring strategy consistently outperforms other methods on test data set, and suggest that the CHAID is a useful classifier for handling low default portfolios with missing value.

Keywords: Credit Scoring; Banking Industry; Rule extraction; Missing data; Low default portfolio.

Received: March 2018-29

Revised: June 2018-14

Accepted: November 2018-26

1. Introduction

Credit scoring is used in banking industry from the past decades and competitive pressures of the industry makes sophisticated scoring models as one of the main assets and sources of competitiveness for banks. It is used to answer one key question - what is the probability of default within a fixed period, usually one year. Data sets of previous loan applicants are built by gathering their data, and these internal data sets are usually completed using external credit bureau's data and finally used for estimating PD in banks, using scoring software's in which the model is in the heart of them (Van Gestel and Baesens, 2009).
There classification techniques in the credit scoring problems include intelligent and statistical techniques. Logistic Regression (LR) is one of the most favorite traditional

* Corresponding author; sadatrasoul@gmail.com
[1] Management school, Kharazmi University, Tehran, Iran.
[2] Shahid Beheshti University, Tehran, Iran.

statistical techniques, and because of its transparency it is used to build credit scorecards(Wiginton 1980). Linear Discriminant Analysis (LDA) is another statistical techniques which is efficient in credit scoring like LR (Harrell and Lee 1985). Intelligent techniques are also applied for credit scoring including rules based and Decision Trees (DT), Bayesian Networks (BN), Neural Networks (NN), Case Based Reasoning (CBR), Support Vector Machines (SVM). SVM, NN, DT and some of intelligent techniques are performing better classification compared to statistical techniques in some studies (Huang, Chen et al., 2004, Ong, Huang et al., 2005, Crook, Edelman et al., 2007). Banks specially the ones which are active in different countries cannot use many of credit scoring techniques specially the intelligent ones because of compliance with transparency, robustness and auditing process done by regulators in some countries in which they operate in (Thomas, 2009). By using rule base classifiers, banks can easily interpret the results and explore the rejecting reasons to the applicant and regulatory auditors and handle their compliance issues with regulators. There literature in the field of rule based credit scoring is little. Rule induction through SVM is also done and showed good results in credit scoring (Martens, Baesens et al., 2007). Adaptive neuro fuzzy inference systems(ANFIS) is also introduced and it's extracted rules works betters than LDA on the studies credit dataset gathered from credit unions(Malhotra and Malhotra 2002), the back propagation is used to learn the rules membership function fitting on the data. A new fuzzy rule induction learning method based on the evolutionary algorithms is also provided and showed better results (Hoffmann, Baesens et al. 2007). A new method is provided for rule pruning and it is examined on the credit scoring data set (Ben-David 2008). Three different NN rule extraction techniques including Nefclass, Trepan and Neurorule is introduced and tested on Bene1 and Bene2 and German credit database. Nero rule and Trepan show better accuracy than C4.5 DT and the LR (Baesens, Setiono et al. 2003).

Missing data could be caused by many circumstances, some due to change in manual application forms design and some to chance, items non-response, partial non-response, previous data aggregation, loss of data, a new established company with no previous data and etc. (Ibrahim, Chen et al., 2005). Low default portfolios (LDPs) characterized by inadequate default records of applicants, because about 85% percent of applicants or even more of them pay back their loans installment's on time. The problem is also escalated when the main missing data are belongs to non-worthy applicants which default. There are three main approaches introduced to handle missing data(King, Honaker et al., 2001, Han, Pei et al., 2011). These strategies are:

- **Eliminate data objects;** it could be a feature or applicant record.
- **Estimate missing values;** based on the feature type estimators like regression, maximum likelihood, C&R tree and etc.; can be used.
- **Ignore the missing value during Analysis;** for the classifiers which can handle the missing values straightforwardly can be used.

The nature of the missing values has main effect on the selection of missing value handling approach. There is also lack of empirical analysis of missing data handling. Listwise Deletion (LD), maximum likelihood (ML) and multiple imputation (MI) are methods which are usually used in credit scoring missing data handling (Florez-Lopez, 2010). Accuracy, robustness and complexity are used to performance checking. MI provides better results in credit scoring problems. Other studies in missing value handling in credit scoring are studied the effect of reject inference in order to estimate PD under the current acceptance policy and it not concern of this paper(Crook and Banasik, 2004, Bücker, van Kampen et al., 2013). There is not a new literature in the field of missing data handling for credit scoring, but data imputation of questionnaires by means of genetic algorithms with different fitness functions

is done (Galán, Lasheras et al., 2017). Random forest missing data algorithms are also applied and shown good results recently (Fei and Hemant, 2017). This paper focused on the of handling missing data in credit scoring data sets in which as mentioned above missing data problem is escalated when the main missing data belongs to non-worthy applicants which default.

The rest of this paper is structured as follows: section 2 describes the rule based classification techniques used. Section 3 introduces the data, main approaches for dealing with missing values and a discussion of their weaknesses and strengths, experiments settings and performance analysis approaches, Section 4 discussed their results and finally study concluded in section 5.

2. Overview of classification techniques

This paper aims to extract the best rules from imbalanced data in the credit scoring context. For this purpose four rule-based and tree induction (with the aim of rule induction) classifiers are selected. Decision trees split the data into smaller subsets using their nodes and at the end of each node there is a series of leaf nodes assigning a class to each of the observations. A rule base can be extracted from each decision tree. Rule bases are more of interest than other classifiers, because they can briefly show why an applicant is rejected and the other one is accepted, however the other classifiers like neural networks, support vector machines cannot lack this ability. A brief description of the rule based classifiers used in this paper is presented below.

2.1. C5

C5 is an extension of C4.5 which mainly used boosting to enhance the results, and it's a favorite model for credit scoring(PANG and GONG, 2009). C4.5 build trees based on the concept of information theory(Quinlan 1993).the entropy of a sample of K, can be computed by:

$$Entropy\ (k) = -\ p_1 log_2(p_1) -\ p_2 log_2(p_2) \quad (1)$$

Where p_1(p_0) are the proportions of the class values 1(0) in the sample K, respectively. The variables which has the highest normalized information gain is picked up for division. The algorithm then occurs on the smaller subsets iteratively.

2.2 CHAID

Chi-square Automatic Interaction Detection (CHAID) is a decision tree based classification technique, which is based on Bonferroni testing. CHAID is frequently applied to data sets with categorized dependent and independent variables. It uses merge and split procedure sequentially based on a chi-square test statistic(Wilkinson, 1992).

2.3 C&R Tree

The Classification and Regression (C&R) Tree node starts by examining the features to find the best split, measured by the reduction in an impurity index that results from the split. The split defines two subgroups (splits are always binary), each of which could subsequently split into two more subgroups in deeper layers, and it continues until one of the stopping criteria is reached(Loh, 2008).

2.4 QUEST

Quick, Unbiased, Efficient Statistical Tree (QUEST) is a binary classification method for building decision trees. A major motivation in its development was to reduce the processing time required for large C&R Tree analyses with either many variables or many cases. A second goal of QUEST was to reduce the tendency found in classification tree methods to favor inputs that allow more splits, that is, continuous (numeric range) input fields or those with many categories(Loh, 2008).

3. Empirical evaluation

In this section the data set characteristics is described first. Then dataset samples which are necessary for the paper experiments are explained and finally the performance analyses metrics are introduced briefly.

3.1. Data sets characteristics

An Iranian bank corporate credit dataset which is used in authors previous studies is also used in this paper, features of the data set is shown in table (9) in appendix (1) (Sadatrasoul, Gholamian et al.).There are few missing values for some corporates, some of them lack financial data and others lack the result of their loans reimbursements and they were in the process of debt repay. 33 features among them (71.7%) have complete data and 813 (81.3%) corporate applicants' data (records) are complete. The dataset characteristics and the changes in the data sets before and after cleaning the data preprocessing descriptions are presented in Table (1). In order to recognizing the datasets better in the research, each one of them are labeled with a data set code which are shown at the first column of table (1).

Table 1. Dataset description

Data set code	Description	Data size	Inputs variables				Complete features %	Complete applicant records %
			Total	Continuous	Categorical	Features with Missing value		
1	Initial dataset	1431	46	38	8	13	NA	NA
2	Dataset (1) with features converted	1431	46	38	8	13	NA	NA
3	431 records from Dataset (2) are eliminated because their loan are current and they are in the process of repay	1000	46	38	8	13	71.7	81.3
4	Data set (3) variable are changed and categorical	1000	54	34	20	13	75.93	81.3

3.2. Missing value handling setup

There are different reasons for missing values, in some cased information is not collected e.g., loan applicants decline to give their data; in some cases attributes may not be applicable to all applicant cases e.g., three prior year annual income is not applicable to the companies

that are established one year ago and etc. Missing values cause some fluctuations on credit scoring models and investigating their effect on the credit scoring models are important; there are three main strategies for missing value handling including:

- Replacing with all possible values (weighted by their probabilities);
- Ignore the missing value during analysis;
- Estimating missing values;
- Eliminating applicant's record with missing value.

In this paper three latter methods are investigated, therefore five missing value handling strategies are considered including:

- Ignoring the applicants with missing values during analysis(I);
- Estimating missing values by the mean of the feature(F);
- Estimating missing values by a random value from the features normal distribution(R);
- Estimating missing values by using C&R tree(C);
- Eliminating applicant's record with missing value (E).

By handling each strategy different datasets are created; Table (2) shows different created datasets for each strategy. Data set number (5) is used for testing the models, because the paper is seeking the missing value, the test data set is selected from dataset number (4)s applicants with complete data record. Other datasets are built using its data audit node by SPSS Clementine 12.0 software tool.

Table 2. Dataset description of different missing value handling strategy

Data set code	Description	Missing value strategy	Data size	Inputs variables				Complete features%	Complete applicant records%
				Total	Continuous	Categorical	Features with Missing value		
5	Dataset (4) nearly 30 percent is selected randomly for test	NA	314	54	34	20	0	100	100
6	Dataset (4) nearly 70 percent is selected randomly for train	I	686	54	34	20	13	75.93	81.3
7	Dataset (5) missing values are replaced with mean value	F	686	54	34	20	0	100	100
8	Dataset (5) missing values are replaced with random value based on normal distribution	R	686	54	34	20	0	100	100
9	Dataset (5) missing values are replaced with values using C&R tree	C	686	54	34	20	0	100	100
10	Dataset (5) 117 applicants with missing values are eliminated	E	559	54	34	20	0	100	100

3.3. Performance analysis

In order to analyzing the results comprehensively, the paper applies two methods for performance analysis. First the Borda count and second the data envelopment analysis (DEA). Each method's input and outputs differs due to the type of analysis and covering their weaknesses. Table (3) shows the performance measures used in each approach.

Table 3. Performance approaches and performance measures used in each one

Performance approaches	Performance measures weights	Accuracy	Number of rules	Number of rules for non-worthy	Average rules length	Number of features	Data size
Borda count	Defined by classifier/missing value strategy	*	*	*	*	*	NA
DEA	Virtually by DEA model	*	*	NA	*	*	*

Five different measures are used to analysis the performance of the constructed rule bases. The performance indicators selected in order to measure the effects of significant difference in number of observations. Classifier accuracy, shows the applicants that are currently classified. Compactness of rules is another issue in rule base systems which measures the illustrative ability of the rule bases; in which at a defined level of accuracy of two sample rule bases, the rule base which has lower number of rules, average rule length and features used is preferred. Finally in the imbalanced data sets, the rules which can discriminate the minority class better are of interest, as the credit data sets have a very small minority class of non-worthy applicants this indicators are very important.

In order to give an overall rank of missing value handling strategies, Borda count is used (Taylor and Pacelli, 2008). The classifier in the first experiments and the missing value strategies in the second experiments are assumed to be voters and the mean vote of them is mentioned for ranking the results.

The paper also applied the data envelopment analysis (DEA) as a comprehensive measure of models efficiency, as it is the method which used in situations that there are not prior weight for input and outputs. Each model is assumed to be a decision-making unit (DMU). The efficiency of each DMU can be evaluated using the generalized DEA estimator (Banker, Charnes et al., 1984).It is assumed that there is constant return to scale (CRS) and therefore the Charnes, Cooper and Rhodes (CCR) estimator is used in the paper (Aldamak and Zolfaghari, 2017). The linear output orientation mathematical model of CCR is used to evaluate the credit scoring models. The credit scoring model accuracy is assumed to be the output and number of rules, average rule length, number of features and number of applicants are assumed to be the input. The efficiency of the models are then solved and reported by solving the CCR output orientation mathematical model 20 times using Lingo 15 software tool.

4. Results and discussions

All the experiments in this paper are done using Table (2) data sets and tests are reported using dataset number (4). Table (4) shows classification accuracy, number of rules, number of rules for non-worthy class label, average rule length and number of features used in the classifier for each model which are labeled with a unique code at the first column of the Table (4). The best classification accuracy, the lowest number of rules, average rule length, number of features used and a few rules for distinguishing non worthy class label is of

interest. It can be seen from table (4) that, there is not a best performer model in all five performance analysis measures. For example although IC5 has the best accuracy it is not better from its competitor models in the other four measures.

Table 4. Performance measures on different missing value handling methods

Model code	Handling missing values strategy/ Type of estimation	Data set	Classifier	Accuracy%	Number of rules	Number of rules for non-worthy	Average rules length	Number of features	DEA efficiency
IC5	Ignore the Missing Value During Analysis	6	C5	87.26	11	5	2.54	10	0.916506785
ICR		6	C&R	85.67	4	2	2.25	18	0.882538676
IQU		6	QUEST	85.99	1	0	0	0	0.997484
ICH		6	CHAID	85.99	17	2	4.53	12	0.877671997
MC5	Estimate Missing Values	mean	7	C5	86.62	9	5	2.56	8	0.927141811
MCR			7	C&R	82.8	6	3	3	19	0.802041512
MQU			7	QUEST	85.99	1	0	0	0	0.997484
MCH			7	CHAID	85.35	15	2	3.6	9	0.896271797
RC5		Random	8	C5	85.03	18	2	3.39	18	0.834089722
RCR			8	C&R	85.67	4	2	2.25	16	0.882538676
RQU			8	QUEST	85.99	1	0	0	0	0.997484
RCH			8	CHAID	85.35	16	2	3.94	12	0.871139725
CC5		C&R	9	C5	83.12	14	6	3.64	18	0.814892178
CCR			9	C&R	82.8	7	3	3.43	24	0.786991922
CQU			9	QUEST	85.99	1	0	0	0	0.997484
CCH			9	CHAID	85.99	15	2	3.73	9	0.902992523
EC5	Eliminate Data Objects	10	C5	83.76	20	10	3.6	20	0.980490089
ECR		10	C&R	85.03	5	2	2.6	18	1
EQU		10	QUEST	85.03	5	1	2.6	19	0.995123664
ECH		10	CHAID	83.12	18	5	4.39	14	1

4.1. Group one experiment (missing value handling strategies performance comparison)

Table (5) shows the rank results against each performance measure for different missing value handling strategies. It can be seen that the mean strategy and ignoring strategy performs better than others. Eliminating strategy is the worst performer in all of measures but it should be taken to consideration that it shows these results with a sample size of 559 compared to others which use 686 samples (about 20% lower sample size).

Table 5. Missing value handling strategies rank against different measures, note: M for mean, I for ignore, R for random, C for C&R tree and finally E for eliminate strategy

Performance measure	Missing value handling strategies rank
Accuracy%	I>R>M>C,E
Number of rules	M>I>C>R>E
Number of rules for non-worthy	R>I>M>C>E
Average rules length	M>I>R>C>E
Number of features	M>I>R>C>E

In order to better comparing the differences, performance measures are shown in a radar chart in figure (1) after standardization. The main difference between the strategies are in number of rules for non-worthy measure, and the accuracy is the most challenging.

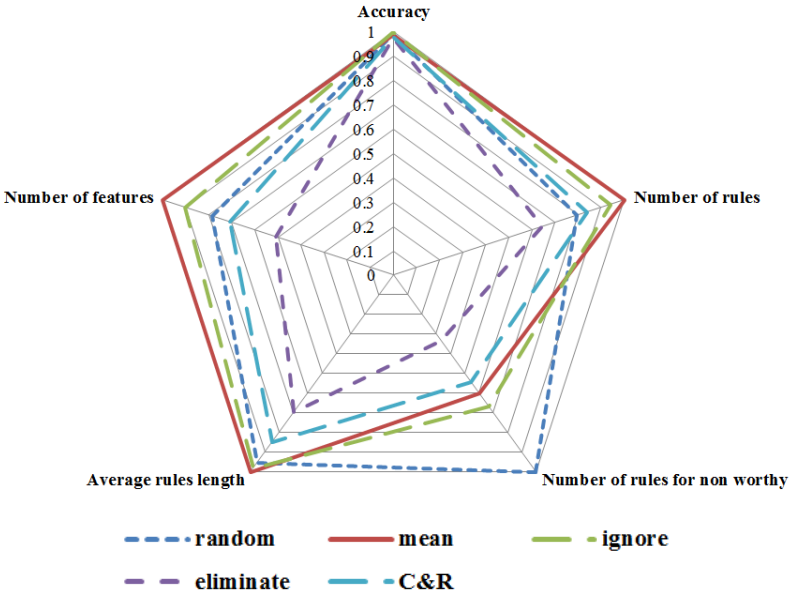


Figure 1. Status of different performance measures for different missing value handling strategies (the results are standardized)

The Borda count is computed using the mean vote of classifiers as the voters and the DEA for each strategy is computed using mean of DEA efficiency of classifiers in a specific strategy. Table (6) shows the overall results of Borda count and DEA efficiency. It can be seen that I>R>C pattern is seen in both approaches. It can be concluded that ignoring is a better strategy than random, and both of them are better than replacing missing values with C&R trees. There is also another extracted pattern M>R>C which also shows that mean is better than random and both are better from C&R trees. Finally the last pattern is I>=M which gives one the analysis that, at the same situation ignoring is preferred to replacing the missing value with mean. Eliminating strategy challenged the analysis also, it is ranked first in one approach and last in another, the main reason for this issue is probably the lower sample size. Although the researchers could take sample balancing techniques for elimination strategy and use techniques like random over sampling, but it is preferred not to use it, because it is unfair for the models competition. The final pattern could be I>=M>R>C.

Table 6. Missing value handling strategies rank against different performance approaches, note: M for mean, I for ignore, R for random, C for C&R tree and finally E for eliminate strategy

Performance approach	Missing value handling strategies rank
Borda count	M,I>R>C>E
DEA efficiency	E>I>M>R>C

4.2. Group two experiments (classifier performance comparison)

Table (7) shows the rank results against each performance measure for different classifiers. It can be seen that QUEST and C&R performs better than others overall. C5 is the worst performer overall.

Table 7. Rule base classifier rank against different measures, Note: C5 for C5, CR for C&R tree, QU for QUEST, CH for CHAID

Performance measure	Missing value handling strategies rank
Accuracy%	QU>CH>C5>CR
Number of rules	CR>C5>CH>QU
Number of rules for non-worthy	QU>CR>CH>C5
Average rules length	QU>CR>C5>CH
Number of features	QU>CR>CH>C5

In order to better comparing the differences, performance measures are shown in a radar chart in figure (2) after standardization. CR is the best performer with significant distance compared to others.

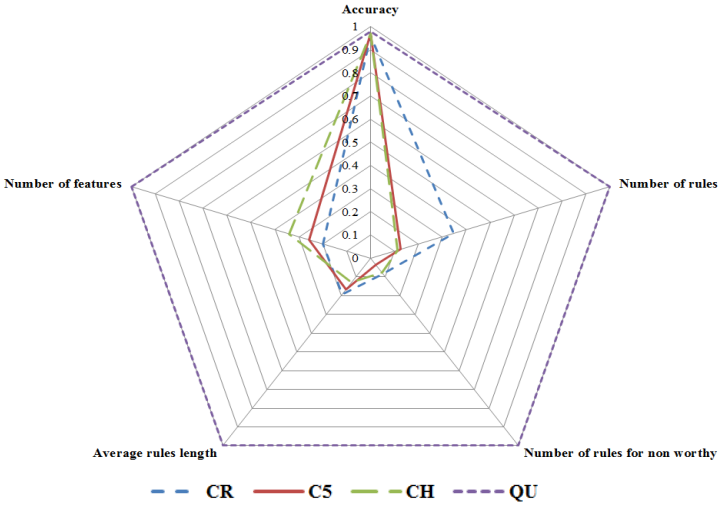


Figure 2. Status of different performance measures for different classifiers (the results are standardized), Note: C5 for C5, CR for C&R tree, QU for QUEST, and CH for CHAID

The Borda count is computed using the mean vote of missing value handling strategies as the voters and the DEA for each strategy is computed using mean of DEA efficiency of missing value handling strategies for a specific classifier. Table (8) shows the overall results of Borda count and DEA efficiency. It can be seen that QU>CH>C5 pattern is seen in both approaches. Therefore QUEST is more robust than CHAID and CHAID is more robust than C5 facing missing data handling confusions. But there is an important issue for verification of this pattern where table (4) results are reconsidered. QUEST use a default creditworthy class and cannot build a real rule set from datasets 6 to 9, in fact it only builds a rule set for database (10), although its mainly because of imbalanced dataset, but for fair competition between models oversampling of non-worthy applicants are not considered. Finally a verified pattern of CH>C5 is acceptable. Another pattern is QU>CR that shows better performance of QUEST to C&R, it also cannot be verified because of the mentioned issue. The final pattern for the classifiers performance could be CH>C5 and unfortunately the research has no consideration for the others.

Table 8. Classifiers rank against different performance approaches, note: M for mean, I for ignore, R for random, C for C&R tree and finally E for eliminate strategy

Performance approach	Rule base classifier rank
Borda count	QU>CR>CH>C5
DEA efficiency	QU>CH>C5>CR

5. Conclusion

In this paper, a number of different rule based classifiers are used and compared in five different missing value handling strategies on a real corporate credit scoring dataset. The missing value handling strategies includes ignoring the missing, replacing the missing with mean value, replacing the missing with a random value base on normal distribution of each feature, replacing the missing with a C&R tree assisted algorithm and finally eliminating the applicants records with missing value. The performance of the strategies as the main research goal are compared using Borda count and DEA, the results shows that ignoring and replacing with mean value are the best strategy and random and C&R tree are orderly the weaker strategies. The performance of the rule based classifiers when facing missing value credit data sets is also investigated as the research's sub goal. The findings show that CHAID is better than C5 totally. Any findings for the eliminating missing value strategy and other rule base classifier could not be explained and next researches can focuses on designing experiments which can brief these two issues.

References

Aldamak, A., and S. Zolfaghari, (2017). "Review of efficiency ranking methods in data envelopment analysis", *Measurement*, Vol. 106, pp. 161-172.

Baesens, B., Sentiono, Rudy, Mues, Christphe, and Vanthienen, Jan, (2003). "Using neural network rule extraction and decision tables for credit-risk evaluation", *Management Science*, Vol. 49, No. 3, pp. 312-329.

Banker, R. D., Charnes, A., and Cooper, W., (1984). "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis", *Management Science*, Vol. 30, Vol. 9, pp. 1078-1092.

Ben-David, A., (2008). "Rule effectiveness in rule-based systems: A credit scoring case study", *Expert Systems with Applications*, Vol. 34, No. 4, pp. 2783-2788.

Bücker, M., (2013). "Reject inference in consumer credit scoring with no ignorable missing data", *Journal of Banking & Finance*, Vol. 37, No. 3, pp. 1040-1045.

Crook, J., and J. Banasik (2004). "Does reject inference really improve the performance of application scoring models?", *Journal of Banking & Finance*, Vol. 28, No. 4, pp. 857-874.

Crook, J. N., (2007). "Recent developments in consumer credit risk assessment", *European Journal of Operational Research*, Vol. 183, No. 3, pp. 1447-1465.

Fei, T., and I. Hemant, (2017). "Random forest missing data algorithms" *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Vol. 10, No. 6, pp. 363-377.

Florez-Lopez, R., (2010). "Effects of missing data in credit risk scoring. A comparative analysis of methods to achieve robustness in the absence of sufficient data", *Journal of the Operational Research Society*, Vol. 61, No. 3, pp. 486-501.

Galán, C. O., (2017). "Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions", *Journal of computational and applied mathematics*, Vol. 311, pp. 704-717.

Han, J., Kamber, and M., Pei, J., (2011). *Data mining: concepts and techniques*, Elsevier.

Harrell, F. E., and K. L. Lee (1985). "A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality", Biostatistics: Statistics in Biomedical; Public Health; and Environmental Sciences. The Bernard G. Greenberg Volume. New York: North-Holland, pp. 333–343.

Hoffmann, F., Baesens, B., Gestel, T., and Vanthienen, J., (2007). "Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms", *European Journal of Operational Research*, Vol. 177, No. 1, pp. 540-555.

Huang, Z., Chen, H., Hsu, C., and Wu, S., (2004). "Credit rating analysis with support vector machines and neural networks: a market comparative study", *Decision support systems*, Vol. 37, No. 4, pp. 543-558.

Ibrahim, J. G., Chen, M., Lipsitz, S., and Herring, A., (2005). "Missing-data methods for generalized linear models: A comparative review", *Journal of the American Statistical Association*, Vol. 100, No. 469, pp. 332-346.

King, G., (2001). "Analyzing incomplete political science data: An alternative algorithm for multiple imputation", American Political Science Association, Cambridge Univ Press.

Loh, W. Y., (2008). "Classification and regression tree methods", Encyclopedia of statistics in quality and reliability.

Malhotra, R., and D. K. Malhotra (2002). "Differentiating between good credits and bad credits using neuro-fuzzy systems", *European Journal of Operational Research*, Vol. 136, No. 1, pp. 190-211.

Martens, D., (2007). "Comprehensible credit scoring models using rule extraction from support vector machines", *European Journal of Operational Research*, Vol. 183, No. 3, pp. 1466-1476.

Ong, C. S., (2005). "Building credit scoring models using genetic programming", *Expert Systems with Applications*, Vol. 29, No. 1, pp. 41-47.

PANG, S. l., and J. z. GONG (2009). "C5. 0 classification algorithm and application on individual credit evaluation of banks", *Systems Engineering-Theory & Practice*, Vol. 29, No. 12, pp. 94-104.

Quinlan, J. R., (1993). *C4. 5: programs for machine learning*, Morgan kaufmann.

Taylor, A. D., and A. M. Pacelli, (2008). *Mathematics and politics: strategy, voting, power, and proof*, Springer Science & Business Media.

Thomas, L. C., (2009). "Consumer credit models: pricing, profit, and portfolios", Oxford University Press, USA.

Van Gestel, T., and B. Baesens, (2009). "Credit risk management: basic concepts: financial risk components, rating analysis, models, economic and regulatory capital", Oxford University Press, USA.

Wiginton, J. C., (1980). "A note on the comparison of logit and discriminant models of consumer credit behavior" *Journal of Financial and Quantitative Analysis*, Vol. 15, No. 3, pp. 757-770.

Wilkinson, L., (1992). "Tree structured data analysis: AID, CHAID and CART", Retrieved February 1: 2008.

This article can be cited: Sadatrasoul, S. M., and Hajimohammadi, Z., (2018). "Investigating the missing data effect on credit scoring rule based models: The case of an Iranian bank", *Journal of Industrial Engineering and Management Studies*, Vol. 5, No. 2, pp. 1-12.

Appendix A.

Table A (1). List of variables in Iran commercial bank credit dataset

Variable	Type	Complete%	Variable	Type	Complete %
Net profit	Continuous	100	industry and mine (=1, other =0)	Categorical	100
Active in internal market	Categorical	100	agricultural (=1, other =0)	Categorical	100
number of countries that the company export to	Categorical	100	oil and petrochemical (=1, other =0)	Categorical	100
Sales Growth	Categorical	97.95	infrastructure and service(=1, other =0)	Categorical	100
Target market risk (from 1 to 5)	Categorical	99.56	chemical (=1, other =0)	Categorical	100
Seasonal Factors	Categorical	100	Year of financial ratio	Continuous	100
Company history(number of years)	Categorical	100	Tax declaration(=1,other=0)	Categorical	100
Top Mangers history	Categorical	100	Audit Organization (=1,other=0)	Categorical	100
Cooperative (=1, other =0)	Categorical	100	Accredited auditor (=1,other=0)	Categorical	100
Stock Exchange (LLP) (=1, other =0)	Categorical	100	Inventory cash	Continuous	100
Generic join stock (PJS) (=1, other =0)	Categorical	100	Accounts receivable	Continuous	100
Limited and others (=1, other =0)	Categorical	100	Other Accounts receivable	Continuous	100
Stock Exchange (=1, other =0)	Categorical	100	Total inventory	Continuous	100
Experience with Bank	Categorical	100	Current assets	Continuous	100
Audit report reliability	binary	93	Non-current assets	Continuous	100
Current period sales	Continuous	100	Total assets	Continuous	100
Prior period sales	Continuous	98.98	Short-term financial liabilities	Continuous	100
Two-Prior period sales	Continuous	97.52	Current liabilities	Continuous	100
Current period assets	Continuous	100	Long-term financial liabilities	Continuous	100
Prior period assets	Continuous	98.83	Non-current liabilities	Continuous	100
Two-Prior period assets	Continuous	98.1	Total liabilities	Continuous	100
Current period shareholder Equity	Continuous	100	Capital	Continuous	100
Prior period shareholder Equity	Continuous	98.68	Accumulated gains or losses	Continuous	100
Two-Prior period shareholder Equity	Continuous	96.94	shareholder Equity	Continuous	100
current accounts creditor turn over	Continuous	99.56	Sale	Continuous	100
Current Account Weighted Average	Continuous	99.41	Gross profit	Continuous	100
Last three years average exports	Continuous	99.56	Financial costs	Continuous	100
Last three years average imports	Continuous	91.98	worthy/unworthy	binary	100