Assessment of Micro Loan Payment using Structured Data Mining Techniques: The Case of Indonesian People' Credit Bank

Novita Ikasari, and Fedja Hadzic

Abstract— **Providing financial service to Micro, Small and Medium Enterprises (MSMEs) in Indonesia presents a challenge for small rural banks such as People' Credit Banks. These banks are required to infer risks about customers' loan repayment from structured (quantitative, financial) and unstructured (qualitative, non-financial) type of credit information. In this study, the complex nature of credit related information is contextualised and represented in domain specific way using the eXtensible Markup Language (XML). An approach that enables the application of wider selections of data mining techniques on XML data is utilized. Experiments are performed using real world credit data obtained from an Indonesian bank. The results demonstrate the potential of the approach to generate reliable and valid patterns useful for evaluation of existing lending policy.**

Index Terms— **Credit risk assessment, Database Structure Model, Data mining techniques, Indonesian People' credit bank, XML**

I. Introduction

THE long debate about outreach and performance as determinants of the success of microfinance institutions has been intensified and documented in well over two decades with unresolved precedence of one over the other. In Indonesia, the issue of outreach has exceeded that of performance especially in the wake of Micro and Small-Medium Enterprises' (MSMEs) achievements in preventing the country's macro-economic indicators from reaching nadir points during global financial crises. In order to provide MSMEs with greater access to financial support, the Indonesian government has taken a proactive stance by bringing financial intermediation closer to those in need, by transforming local financial institutions into a well-structured and legalised business called People' Credit Bank (*Bank Perkreditan Rakyat).* Through Law Number 10 Year 1998 on Banking, people' credit banks are formalised as one type of banks in Indonesia with limited banking services.

Targeting at MSMEs, with more emphasis on micro business, people' credit banks provide straightforward savings and lending services to their customers. Since these banks are located in the heart of business in one region, they are adapted to satisfy customers' specific business needs and demands of loan services. This requires banks to carefully develop and assess industry-sensitive credit risk profile to ensure timely and full loan repayment. In general, banks are facing the issue of asymmetric information when delivering financial services to MSMEs as a result of MSMEs' business management style [1]. The need to make an informed lending decision for MSMEs has become more crucial to people' credit banks since their business is operated on small capitalisation of less than IDR 100 billion. The standardised credit risk assessment as regulated by the Central Bank, known as 5Cs good lending concept, incorporates information on borrower' *c*haracter, *c*apacity (to make payments), *c*apital, *c*ollateral and relevant *c*onditions of economy. It is a challenging task for people' credit banks to acquire valid and reliable information on prospective borrowers' 5Cs. It is even more taxing to perform an assessment on such information provided that banks are flooded with structured (quantitative, financial) and unstructured (qualitative, non-financial) type of data. In contrast with lending for big-sized companies, this information is scarce, unreliable and needs to be verified through various secondary sources as MSMEs are operating in a much less structured manner than their big-sized companies counterparts.

With regards to credit risk assessment methods, a considerable amount of literature has pivoted around credit scoring, covering advantages such as efficient decision making [2], [3] and more reliable outcome [4], [5]. In addition, existing literature has documented the dominant role of structured financial information [6], [7] and the supplementary role of structured non-financial information [2], [8] to arrive at loan granting decision for small businesses. Unfortunately, available methods are yet to be well-established to provide satisfactory credit risk assessment methods for MSMEs that incorporates structured (categorical) and unstructured (text) types of data. We have commenced an effort to present structured and unstructured data in one template and at the same time preserve the context in which this information resides by using eXtensible Markup Language (XML) as detailed in [9]. Nevertheless, the data has become extremely complex due to the existence of contextual nodes that are required to

Manuscript received March 16, 2012; revised April 11, 2012.

Novita Ikasari is with School of Economics and Finance, Curtin Business School, Curtin University, Perth, Australia and Faculty of Social and Political Science, University of Indonesia, Depok, Indonesia (Phone: 61-8-9266-9275; fax: 61-8-9266-7548; e-mail: novita.ikasari@postgrad.curtin.edu.au).

Fedja Hadzic is with Department of Computing, Faculty of Science and Engineering, Curtin University, Perth, Australia (e-mail: fedja.hadzic@curtin.edu.au).

preserve the hierarchy and contextual meaning of the information as will be elaborated in subsequent section. Based on our previous work, this study is set to contribute to MSMEs' credit risk assessment in twofold. First, it attempts to fill the existing gap in literature on effective approach to discover the body of knowledge from structured and unstructured data. Second, we aim to provide practical support to the banking industry by furnishing The Bank with determinants necessary to understand customers' loan payments that subsequently would benefit MSMEs lending strategy' formulation. Experiments are performed using credit data provided by our industry partner (hereafter The Bank). A structure-preserving flat data format of tree-structured data such as XML has been recently proposed in [10]. The main motivation is to enable a wider range of data analysis/mining techniques to be directly applied on tree-structured data, and to alleviate the complexity associated with mining complex structures. We explore this technique in this study and use association rule mining and decision tree learning methods as case in point. The knowledge patterns obtained through different data mining techniques are discussed for their implications in the domain.

The remainder of this paper is structured as follows. In the next section we describe the credit risk assessment method applied by The Bank. In Section III the parallelism between XML and tree-structure is illustrated. This lays the necessary ground for understanding the applied method discussed in Section IV. Section V presents the experiments and discussions of findings related to credit policy revision. The paper is concluded in Section VI.

II. Five Cs' good lending concept and Data profile

It is economically rational for The Bank to exercise 5Cs in a more straightforward manner where information on each of the Cs is captured irrespective of the designated label. In this section we highlight the 5Cs as practiced by The Bank and the profile of credit data.

A. *Implementation of 5Cs*

The first C, "**Character**", refers to The Bank's ability to establish customers' willingness to pay. Provided the quite wide recognition of the surrounding community, many of The Bank's customers are recurring debtors with loan history being systematically and manually recorded in records' storage. Within this context, The Bank practice relationship lending where debtors have built professional and personal rapport with loan officers. Since many customers are local inhabitants who run the business at his/her home, the second C, **"Capacity (to pay)"** that takes a form of a simple but informative financial report, is constructed from interviews and observations. Next, **"Capital"** is simplified by taking the amount of cost of goods sold as the company's capital proxy. The safety net for banks, **"Collateral"** is categorized into four types, namely land and building, motorcycles, cars and bank's deposit. The last C, **"Conditions (of economy)"**, is constructed in an inward looking manner whereby The Bank requires information on how debtor oversees the chain of supply within industry competition and services.

B. *Credit Data*

The total number of loan applications used for this study is 96 records with 58 performing and 38 non-performing loans at the time of collection. The area in which The Bank is located is renowned for its density and small business establishments. It is also customary for small entrepreneurs to have more than one business when he/she is the sole owner of such venture, which is the case with 35 debtors (36%) in our dataset. These businesses are not necessarily supplementary or even related to the main one. Additional ventures denote both positive and negative notion to banks. They are considered a strong point for borrower as it strengthens borrower's professional portfolio and increases the likelihood of meeting his/her loan obligations on time as a result of the existent of extra source of income. However, the lack of focus in business also implicates higher demand on resources and financing. Majority of local community is traders with amount of loan principle concentrated on the low to middle range (IDR 1 million up to IDR 50 million).

As part of data pre-processing, the numerical values were converted into defined categories, using discretization methods [11]. This is necessary to reduce the number of unique values of an attribute and detect similarities (form generalizations) during the data mining process, in spite of subtle differences between original numerical values. In performing assessment on loan application, The Bank categorized borrower's credit information into Objective and Subjective Analysis in order to have straightforward understanding of borrower's business.

Objective Analysis contains factual numeric information on cash inflow and outflow incurred by the business within one financial year. Subjective Analysis contains supplementary business as well as non-business related information in text format. Other than these, The Bank collects information on loan structure such as loan principal, interest rates, and purpose of loan, as well as on collateral such as type, and value of collateral. With almost 80% of the attributes containing specific numeric attribute values, discretization becomes essential. Discretization on attributes that is displayed on a separate part of the loan application document, such as loan principal, loan duration and type of collateral, is done by domain expert in accordance with internal directive memorandum.

Data preprocessing is carried out on Subjective Analysis by manually identifying implicit structure within the text of each sub section. The first section, "Existing Business Standing" is very structured with categories such as "growing", "stabile" or "slowing down". The structure in information on the other three sections, which are "networking with buyers", "networking with suppliers" and "management style", inherently exist since the Bank needs to maintain standardized interview questions.

C. *XML Document*

In domains where the nature of the data is more complex and a domain-specific way of organizing the available data is required, semi-structured documents such as XML are often used [12]. Our motivation to use XML is to capture both types of data (structured and unstructured) in a domain-specific way and effectively organize (contextualise) the available information. Hence, following data preprocessing, credit data is then populated into an XML document based on a pre-defined XML template. Fig. 1 shows one credit application of the resulting XML document with selected attributes (subset) and values.

```
<?xmlversion="1.0"encoding="utf-8"?>
<CreditApplication id="PSP01">
    <loanapplication>
        <debtorstatus>recurrent</debtorstatus>
        <industry1>trade</industry1>
        <industry2>nr</industry2>
        <industry3>nr</industry3>
        <loanscheme>
            <principal>[100000000-249999000]</principal>
            <dailyprincipal>[140277.7778 -324074.0741]</dailyprincipal>
            <dailyinstallment>[225000-277370] </dailyinstallment>
            <percentageofdailyinstallment>[36.0 - 36.08]
            </percentageofdailyinstallment>
            <dailyinstallmentdeposit>[4500-23580]
            </dailyinstallmentdeposit>
            <percentageofdailyinstallmentdeposit>[60.0 - 61.0]
            </percentageofdailyinstallmentdeposit>
            <dailyinstallmentanddeposit>[315000.0 - 540000.0]
            </dailyinstallmentanddeposit>
            <dailiyinstallanddeposittodailyloan>[1.962 -
            2.0829]</dailiyinstallanddeposittodailyloan>
            <loanduration>720</loanduration>
            <interestrate>[14.4 - 19.0]</interestrate>
        </loanscheme>
    </loanapplication>
    <creditperformance>performing</creditperformance>
</CreditApplication>
...
</xml>
```

Fig. 1. A fragment of XML template

III. XML and Tree Parallelism Illustration

An XML document has a hierarchical document structure, where an element may contain further embedded elements, and each element can be attached with a number of attributes. It is therefore frequently modeled using a rooted ordered labeled tree. Several works [13] - [15] have been proposed for mining of semi-structured documents such as XML. Initially, the focus was mainly on values associated with the tags, which is by and large no different from traditional association rule mining. However, for certain application domains and to maximize the information content of discovered knowledge patterns, it is necessary to take the structural information of the document into account. In the remainder of this section, definitions and concepts have been reproduced and readapted from [12].

A. Tree Concepts and Definitions

A graph contains a set of nodes (or vertices) connected by edges, and each edge has two nodes associated with it. A path is a finite sequence of edges, and it length equals to the number of edges. A rooted tree has its top-most node that has no incoming edges defined as the root. In a tree a single unique path connects any two nodes. If there is a directed edge from *u* to *v*, node *u* is said to be a parent of node *v*, while node *v* is a child of node *u*. Sibling nodes share the same parent. The number of children of a node defines the fan-out/degree of that node. The level/depth of a node is the length of the path from root node to that node. The height of a tree corresponds to the largest level of its nodes.

B. XML Document Entities

Nodes can be categorized into simple and complex nodes [14]. Simple or basic nodes have no edges emanating from them. In a tree structure, this type of node is called a leaf node. Complex nodes are also called internal nodes. Two important relationships can be constructed from complex nodes: parent-child and ancestor-descendant. These two relationships are equivalent to the same parent-child and ancestor-descendant relationships defined in previous section and mark the important parallelism between XML and tree structure. From Fig. 2, representative simple nodes examples (from first credit application) would be <industry1>, <industry2>, <industry3>, <principal>, <dailyprincipal>, <dailyinstallment>, etc. The complex nodes examples are <CreditApplication>, <loanapplication>, <loanscheme>.

The element-attribute relationship in XML is of significant value. When it comes to tree structure, this is more or less a depiction of a node with multi-labels and the level of relationships among them is of equal value. One is no more important than any other. Relationships between elements in XML are the basic construct for hierarchical relationships. The relationships of two elements are either parent-child relationships or ancestor-descendant relationships. The two elements that are connected by one edge are the basis for a parent-child relationship. The two elements that are connected by more than one edge are the basis for ancestor-descendant relationships. The element-element relationship must be constructed only by two elements from different levels. If two elements have the same levels and belong to the same parent, the relationship between them is a sibling relationship that has no edges connecting them, so it is more of a virtual relationship. Examples of both element-element and element-attribute relationships are shown in Fig. 2. The element-element relationship is that of parent-child, while the element-attribute relationship is that of ancestor-descendant.

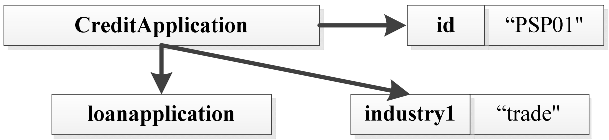

Fig. 2. Illustration of element-element (top to bottom) and element-attribute relationships (top to right)

C. Tree-structured items

The basic construct of XML is tree-structured items. XML contains more complicated hierarchical relationships between tree-structured items, than there exist with relational data. Examples of tree-structured items from Fig. 1 are shown below in Fig. 3.

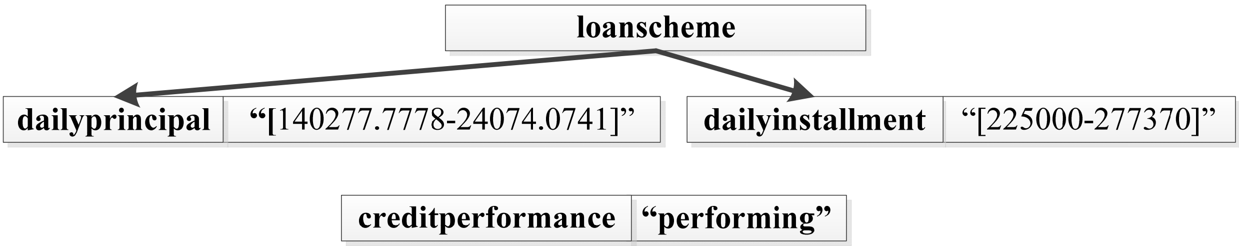

Fig. 3. Illustrations of tree-structured items with size 3 (top) and with size 1 (bottom)

Hence, an XML document has a hierarchical document structure, where an XML element may contain further embedded elements, and these can be attached with a number of attributes. Elements that form sibling relationships may have ordering imposed on them. Each element of an XML document has a *name* (e.g. dailyprincipal) and can have a *value* (e.g. "[140277.7778-324074.0741]"). Given such parallelisms, an XML document can therefore be modeled as a rooted ordered labeled tree. A rooted ordered labeled tree can be denoted as $T(v_0,V,L,E)$, where (1) $v_0 \in V$ is the root vertex; (2) V is the set of vertices or nodes; (3) L is the set of labels of vertices, for any vertex $v \in V$, $L(v)$ is the label of v; and (4) $E = \{(x,y)|x,y \in V\}$ is the set of edges in the tree.

If structures and values are to be considered, XML can be transformed into a string such that the elements' names and their inner text (value) are appended as a single string label (i.e. node/vertex label $L(v)$). From Fig. 2, the element <industry1> can be transformed into a single string 'industry1["trade"]'. The data representation/mining approach adopted in this work and described in [10] utilizes the pre-order (depth-first) string encoding [15]. Since the processing of long strings can be very expensive, a common strategy used to expedite the processing of XML documents is to transform the string encoding into an integer-based form [10]. With this approach, the textual content of each element node will be mapped into an integer number and the mapping is stored in an index table where the original string can be looked up at later time for reporting purposes.

TABLE I
Example of String to integer mapping for XML fragment of Fig. 1

CreditApplication	0
loanapplication	1
debtorstatus["recurrent"]	2
industry1["trade"]	3
industry2["nr"]	4
industry3["nr"]	5
loanscheme	6
principal["[100000000-249999000]"]	7
dailyprincipal["[140277.7778 - 324074.0741]"]	8
dailyinstallment["[225000-277370]"]	9
percentageofdailyinstallment ["[4500-23580]"]	10
dailyinstallmentdeposit ["[4500-23580]"]	11
percentageofdailyinstallmentdeposit ["[60.0 - 61.0"]	12
dailyinstallmentanddeposit ["[315000.0 - 540000.0]"]	13
dailyinstallanddeposittodailyloan ["[1.962 - 2.0829]"]	14
loanduration ["val_720.0"]	15
interestrate ["[14.4 - 19.0]"]	16
creditperformance["performing"]	17

Table 1 is an example of a mapping between the strings (elements (and values)) and unique integers for the XML fragment in Fig. 1. One can use any hash function to map such strings into integer indexes. With this string to integer mapping, the pre-order string encoding representation of the underlying tree structure of the example credit application of Fig. 1 is transformed to "0 1 2 -1 3 -1 4 -1 5 -1 6 7 -1 8 -1 9 -1 10 -1 11 -1 12 -1 13 -1 14 -1 15 -1 16 -1 -1 -1 17" with corresponding tree shown in Fig. 4.

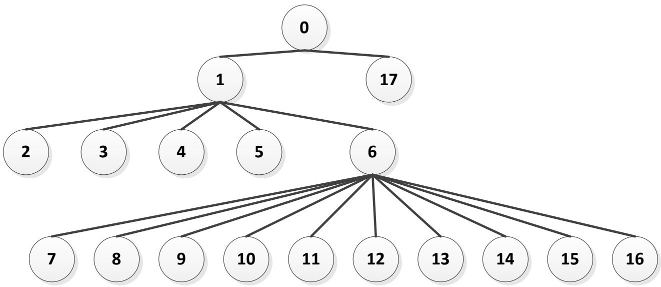

Fig. 4. Integer-indexed tree of XML fragment of Fig. 1

I. Method

As mentioned in the introduction, our previous work in mining of credit application data represented in XML format was limited to the analysis as enabled by frequent subtree mining algorithms. There was combinatorial complexity arose from attributes that are present in every instance or transaction merely to contextualize the available information and not useful for discriminating the class attributes. Therefore, there is a need to develop alternative data mining methods for tree-structured data to directly mine for discriminative patterns with respect to the classification task. In our previous work [9], when an algorithm for mining frequent closed/maximal subtrees [16] was applied, the support threshold required to extract the underlying rules had to be so low at which the algorithm failed to return the results. In addition, at low support thresholds, the frequent subtree patterns are often excessive, causing significant delays in the analysis and interpretation of the results.

A new method for effectively representing tree-structured data into a structure preserving flat format proposed in [10] has the main motivation of enabling a wider range of well-established data mining/analysis techniques, previously developed for flat data format, to be applied directly to tree-structured data. It is promising in the sense that many of the complexity issues caused by the structural properties in the document can be overcome, and class distinguishing criteria can be directly sought after. The technique converts the string like representation commonly used by frequent subtree mining into a flat data structure format (henceforth referred as table) so that both structural and attribute-value information is preserved. The first row of a (relational) table consists of attribute names, which in a tree database are scattered through independent tree instances (credit applications in our case). One way to approach this problem is to first assume a structure according to which all the instances/transactions are organized. Each of the transactions in a tree-structured document should be a valid subtree of this assumed structure, referred to as the *database structure model* (*DSM*) in [10]. This *DSM* will become the first row of the table, and while it does not contain the attribute names, it contains the most general structure where every instance from the tree database can be matched to. The *DSM* needs to ensure that when the labels of a particular transaction from the tree database are processed, they are placed in the correct column, corresponding to the position in the *DSM* where this label was matched to. The string encoding is used to represent the *DSM* and since the order of the nodes (and backtracks ('-1')) is important the nodes and backtracks are labeled sequentially according to their occurrence in the string encoding. For nodes (labels in the string encoding), x_i is used as the attribute name, where *i* corresponds to the pre-order position of the node in the tree, while for backtracks, b_j is used as the attribute name, where *j* corresponds to the backtrack number in the string encoding.

In our example, from Fig. 4, the string encoding of *DSM* becomes 'x_0 x_1 x_2 b_0 x_3 b_1 x_4 b_2 x_5 b_3 x_6 x_7 b_4 x_8 b_5 x_9 b_6 x_{10} b_7 x_{11} b_8 x_{12} b_9 x_{13} b_{10} x_{14} b_{11} x_{15} b_{12} x_{16} b_{13} b_{14} b_{15} x_{17}'. This becomes the first row in the table and to fill in the remaining rows the string encoding [18] of each record is traversed from the tree database. When a label is encountered, it is placed to the matching column under matching node (*x_i)* in the *DSM* structure. When a backtrack ('-1') is encountered, a value '1' is placed to the matching backtrack (b_j). Remaining entries are assigned a value of '0' (non-existence). All instances are organized using a pre-defined XML template, thus it is unnecessary to store the backtrack attributes (-1 or b_j) in the populated table. To indicate structural characteristics of the knowledge patterns discovered, they can be re-mapped to the *DSM.* Hence, structural complexity is avoided and structural characteristics of the data are preserved. For pseudo code of the conversion process and illustrative examples, please refer to [10].

II. Experimental Findings

The structural characteristics of the XML data are as follows: 701 unique labels (attributes and their values when applicable), 82 nodes in each tree instance (transaction), and the height and fan-out of each tree instance were 4 and 12 respectively. The conversion approach described in Section IV produces a standard (relational) table format of this XML data, consisting of a total of 82 attributes (as 81 backtrack attributes (b_j) are omitted but kept in *DSM* to preserve the structure), each mapping to a particular element (and value if that element has one) from the original XML document. The class to be predicted is "creditperformance" with possible values of "performing" and "non-performing". Using the structure-preserving flat format representation [10] of the data a wider range of data analysis/mining techniques can be applied. We have used the association rule mining and decision tree learning algorithms from the publicly available machine learning workbench Weka [17] as a case in point.

A. *Association Rule Mining*

The Apriori algorithm is applied to generate association rules [18]. The Apriori is used on 70% of the data (training set) with minimum confidence of 90% and 10% support. It generated 116,654 rules which were evaluated for their classification accuracy on the same dataset and predictive accuracy using the "unseen" 30% of the data (testing set). The rule set had high accuracy level on both training (98.03%) and testing (89.71%) data with 100% coverage rate. Below are two examples of rule representing good and bad case of loan payment. For the first rule we also provide an example of its subtree based representation in Fig. 5 when nodes are matched to the extracted *DSM* [10].

totalinstallmentanddepositperday([17500-149000]) AND customer(local) AND paymentrisklevel(low)→ creditperformance(performing) (7)

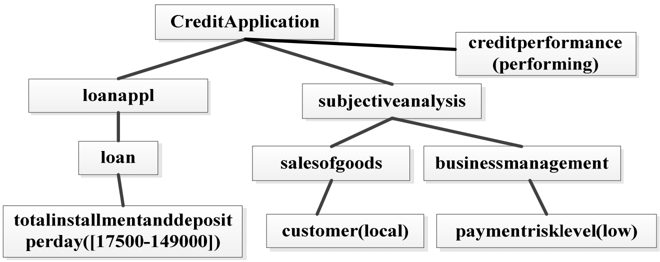

Fig. 5. Subtree based representation of first rule with contextual nodes indicated

The striking attribute on the first rule is the type of customer which raises the locality issue with local customer being regarded as having good performance in loan repayment. The highlight on locality could be argued as an influence from culture. With the vast majority of Indonesian population being a Moslem, Eid is one of the – if not the most – celebrated religious holiday where people go to their home land to be with their extended family. If customer is an immigrant, during this time, business activities are being put on hold and will resume back to their normal pace in a fortnight time. The Bank has a policy regarding a failure-to-pay whereby borrower is required to pay the sum of the non-payment days at their first appearance. Therefore, The Bank preferred to lend money to local people rather than to immigrants. Within this context, The Bank needs to be mindful on taking cultural aspect into account to avoid being discriminative in providing financial service.

purposeofloan(additional capital) AND typeofcollateral(vehicle) AND totalinstallmentanddepositperday([17500-149000]) AND character(no information) → creditperformance(nonperforming) (7)

Purpose of loan and characters of borrowers come into the equation and, as anticipated, influence loan performance in a proportional manner. With regards to the type of collateral given to The Bank, it is confirmed through an interview with the owner of the bank that a moving asset is less preferred to be submitted as collateral than a permanently built and secured type of assets such as land and buildings. The Bank experienced quite a few negative experiences where they are not able to acquire the vehicle when loan payment deteriorated due to a “misplacement” of vehicle or a result of illegal transfer of rights to someone else. One striking fact according to this pattern is that the lowest amount of daily installment and deposit is deemed to lead to a delayed loan. This is something that The Bank should be concerned and acted upon since it contradicts common perception of MSME lending, in particular micro lending, where micro borrowers are renowned for their daily loan payment track record. The argument behind this abnormality is more of ease of procedure rather than ill-intention. The pattern brings about an issue of price-floor, and the necessity to apply this to MSME lending. Small business entrepreneurs are also notorious for their financial mismanagement whereby business’ liquid resources are used to finance non-business activities, hence the importance of collecting information on non-operating expenses such as household expenses, children’ education and health expenses, etc. It is possible that given the extremely low amount of money to be paid to The Bank, both parties (borrowers and The Bank) have adopted relaxed and flexible approach under the assumption that it is easily recovered in the next payment.

B. Decision Tree Learning

Decision tree learning methods have had popular use in the credit risk assessment domain, since the underlying rules are easily interpreted and the method gives good accuracy results. The C4.5 decision tree algorithm [19] was used in this experiment. We have intentionally excluded sub-industry 2 to 5 since The Bank is not expected to have control on the choice of business of customers at the time of loan application.

The C4.5 algorithm generates a decision tree of size 77 with 68 leaves (rules) in 0.01 seconds; with 94.34% accuracy evaluated using 10-fold cross-validation. The parent nodes of this tree are “otherinstallmentexpense (nr)”, “otherinstallmentexpense ([3200000.0 - 3333500.0])”, “otherinstallmentexpense ([417000.0 - 1333500.0])”and “otherinstallmentexpense (val_9300000.0)”. Other installment expense is ranged from as typical as installment on vehicle, which is insignificant in amount, to something as risky as loan payment in other bank(s). As can be seen from the parent nodes, this attribute is categorized into 4 groups. Of these groups, only the group without any attribute value is leading to some of performing loans, which implies the importance of customers’ business and cash management. The rest of the groups have entirely resulted in non-performing loans. Below are two example rules from the group without other loan obligations but with different outcome.

otherinstallmentexpense (nr) AND riskonpayment(low) AND securityandcleaningexpense ([75000.0 - 100000.0]) AND dailyinstallmentdeposit([23600-42680]) → creditperformance(performing) (2)

otherinstallmentexpense (nr) AND riskonpayment(low) AND securityandcleaningexpense ([75000.0 - 100000.0]) AND dailyinstallmentdeposit ([4500-23580]) → creditperformance(nonperforming) (11|2)

From above examples, the only distinct attribute that set good customers from bad ones is the amount of daily installment deposit that the customer has to pay to The Bank. Regardless of the fact that customers do not have other loan obligation to third party, considered to have low risk of delayed payment and have to compensate the security and cleaning service with the same amount of money, customers with the lowest payment obligation (IDR 4,500 to IDR 23,580) have posed higher risk than those who have highest payment obligation. Nevertheless, while this rule correctly classified 11 cases, 2 cases were incorrectly classified. Following these findings, The Bank is best advised to carefully assess customers’ payment capacity, especially those of micro borrowers, before assigning the amount of daily installment deposit.

III. Conclusion

In this paper, we present a way to automatically analyze approved MSMEs’ loan applications to distinguish characteristics that indicate performing and non-performing loans. The quantitative and qualitative data obtained from the industry partner are combined and structurally represented in a domain-specific way using an XML template. A suitable technique is utilized that can handle the complex nature of credit data by directly arriving at class discriminating factors. The experiments demonstrate that the approach can yield granular rules that contribute to academic and practical field of credit risk assessment. Some interesting rules related to micro borrowers have been presented which can be used to provide evidence-based decision support for internal policy refinement. In our future work, we will advance our analysis to discover knowledge patterns targeted towards predicting the periods of delayed payment.

References

[1] A.N. Berger, L.F. Klapper and G.F. Udell, “The Ability of Banks to Lend to Informationally Opaque Small Businesses”, *Journal of Banking & Finance*, vol. 25, pp 2127-2167, 2001.

[2] T.H.T. Dinh and S. Kleimeier, “A Credit Scoring Model for Vietnam's Retail Banking Market”, *International Review of Financial Analysis,* vol. 16, pp 471-495, 2007.

[3] W.S. Frame, A. Srinivasan and L. Woosley. “The Effect of Credit Scoring on Small-Business Lending”, *Journal of Money, Credit, and Banking,* vol. 33, pp 813-825, 2001.

[4] H. Abdou, J. Pointon and A. El-Masry, “Neural Nets versus Conventional Techniques in Credit Scoring in Egyptian Banking”, *Expert Systems with Applications,* vol. 35, pp 1275-1292, 2008.

[5] K.H. Chye, T.W. Chin and G.C. Peng, “Credit Scoring Using Data Mining Techniques”, *Singapore Management Review*, vol. 26, pp 25-47, 2004.

[6] R. Tsaih, , Y.-J. Liu, W. Liu and Y.-L. Lien, “Credit Scoring System for Small Business Loans”, *Decision Support Systems,* vol. 38, pp 91-99, 2004.

[7] C. Wu and X.-M.Wang, “A Neural Network Approach for Analyzing Small Business Lending Decisions”, *Journal Review of Quantitative Finance and Accounting*, vol. 15, pp 259-276, 2000.

[8] B. Lehmann, “Is It Worth the While? The Relevance of Qualitative Information in Credit Rating”, SSRN eLibrary, 2003.

[9] N. Ikasari, F. Hadzic and T.S. Dillon, “Incorporating Qualitative Information for Credit Risk Assessment through Frequent Subtree Mining for XML”, in *XML Data Mining: Models, Method, and Applications*, 1st ed. A. Tagarelli, Ed. USA: IGI Global, 2012, pp 467-503.

[10] F. Hadzic, "A Structure Preserving Flat Data Format Representation for tree-Structured Data", in Proc. 2nd Workshop on Quality Issues, measure of interestingness and evaluation of data mining models, *LNCS/LNAI Post Proceedings of PAKDD Workshops, Springer*, China, May 2011.

[11] J. Han and M. Kamber, "Data Mining Concepts and Techniques", 2nd ed., CA: Morgan Kaufmann Publishers, 2006.

[12] F. Hadzic, H. Tan and T.S. Dillon, "Mining of Data with Complex Structures", *Studies in Computational Intelligence Series*, vol. 333, Springer, Berlin/Heidelberg, Germany, 2011.

[13] Y. Chi, S. Nijssen, R.R. Muntz and J.N. Kok, "Frequent Subtree Mining - An Overview", *Fundamenta Informaticae, Special Issue on Graph and Tree Mining*, vol. 66, no.1-2, pp 161-198, 2005.

[14] K. Wang and H. Liu, "Discovering typical structures of documents: a road map approach", in Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, , Australia, 24-28 August 1998. pp 146-154.

[15] M.J. Zaki, "Efficiently mining frequent trees in a forest: algorithms and applications", *IEEE Transaction Knowl Data Engineering*, vol.17, no. 8, pp. 1021-1035, 2005.

[16] Y. Chi, Y. Yang, Y. Xia and R.R. Muntz , "CMTreeMiner: Mining both closed and maximal frequent subtrees", in Proc. The Eight Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD'04), Australia, May 2004, pp 63-73.

[17] G. Holmes, A. Donkin and I.H. Witten, "Weka: A machine learning workbench", in Proc. 2nd Australia and New Zealand Intelligent Information Systems Conference, Australia, 29 November – 2 December 1994, pp. 357 – 361.

[18] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases", in Proc. 20th International Conference on Very Large Data Bases (VLDB), Chile, 12-15 September 1994, pp. 487-499.

[19] R. Quinlan, *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmann Publishers, 1993.