DE GRUYTER OPEN

Decision Tree Approach to Discovering Fraud in Leasing Agreements

Ivan Horvat
VB Leasing d.o.o., Croatia
Mirjana Pejić Bach
Faculty of Economics & Business – Zagreb, University of Zagreb, Croatia
Marjana Merkač Skok
Fakulteta za poslovne in komercijalne vede, Slovenia

Abstract

Background: Fraud attempts create large losses for financing subjects in modern economies. At the same time, leasing agreements have become more and more popular as a means of financing objects such as machinery and vehicles, but are more vulnerable to fraud attempts. **Objectives:** The goal of the paper is to estimate the usability of the data mining approach in discovering fraud in leasing agreements. **Methods/Approach:** Real-world data from one Croatian leasing firm was used for creating tow models for fraud detection in leasing. The decision tree method was used for creating a classification model, and the CHAID algorithm was deployed. **Results:** The decision tree model has indicated that the object of the leasing agreement had the strongest impact on the probability of fraud. **Conclusions:** In order to enhance the probability of the developed model, it would be necessary to develop software that would enable automated, quick and transparent retrieval of data from the system, processing according to the rules and displaying the results in multiple categories.

Keywords: decision tree; fraud detection; leasing fraud; cars; data mining; leasing agreements

JEL main category: G
JEL classification: G32, O31
Paper type: Research paper

Received: 21, September, 2013
Accepted: 28, March, 2014

Citation: Horvat, I., Pejić Bach, M., Merkač Skok, M. (2014), Decision tree approach to discovering fraud in leasing agreements, Business Systems Research, Vol. 5, No. 2., pp. 61-71
DOI: 10.2478/bsrj-2014-0010

Introduction

Leasing is a modern financing method developed in the U.S.A. in the 30s of the last century, and has been widely accepted and applied in the world from 1950s onwards. Leasing allows the user to use needed equipment or property for a

required period of time, rather than to buy it. A leasing object is a movable or an immovable thing in accordance with the applicable rules governing property or other proprietary rights (Smith, Wakeman, 1985; Morais, 2013).

A leasing agreement becomes realized and active after being signed by a leasing company and a customer. There is no delay in activation or conditional activation of the agreement. There are two main ways in which a leasing agreement can be terminated: the expiration of the agreement and the premature termination. The circumstances that lead to an early termination can be divided into the circumstances caused by users of the lease (total loss, failure to pay monthly installments) and the circumstances caused by external influences (theft, total loss due to natural disasters).

If the agreement is terminated and the attempt to perpetrate fraud or deception is found, the damage for a leasing house is created. Therefore, risk management and using credit scoring are important levers for increasing the security of a leasing company. Advanced analytical methods of assessing the risk of fraud have proved successful in predicting one of the two possible outcomes of the agreements: a successful implementation and finalization of the agreement and an attempted fraud (Ngai et al., 2011; Bhattacharyya et al., 2011; Huang et al., 2012). However, in previous studies, leasing has not been the subject of modeling knowledge discovery from databases, although the method is often used in practice. Therefore, the aim of the paper is to develop a model for detecting fraud in the lease, using actual data from a leasing company. To achieve the objective, knowledge discovery from databases was used and the decision tree method was applied (Sinha, Zhao, 2008).

Methodology

Data

The used database contains information on all leasing agreements and offers in the core system on the date of running the report. The number of active or completed agreements at the time of running the report was 25,000. In the same period a total of 561 agreements in which fraud was realized were found. In order to ensure the possibility of forming a decision tree model, the method of under sampling was used and 560 agreements with no fraud attempts were randomly selected from the total number of observed agreements.

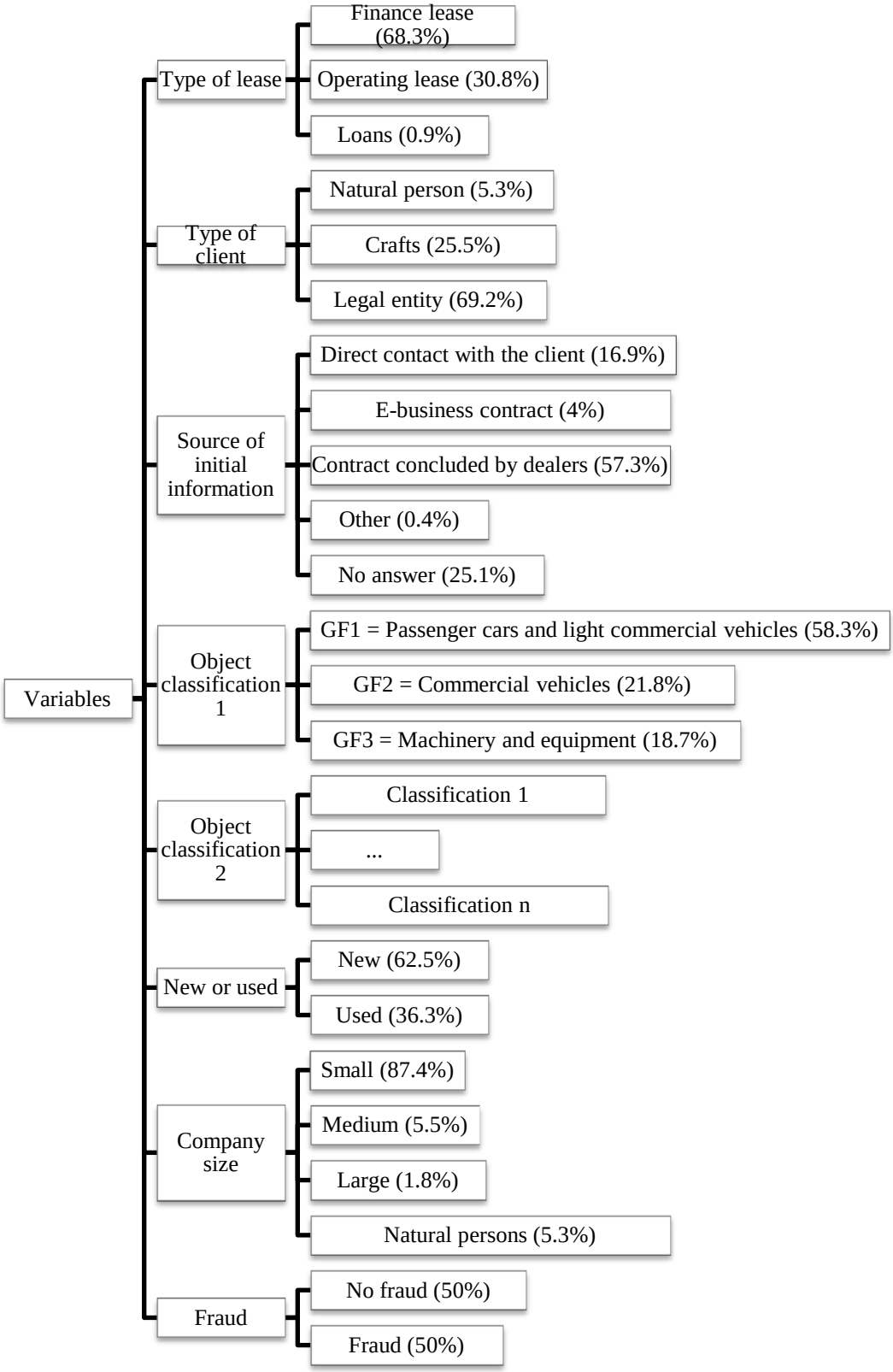
Although the database contains more than a hundred variables, due to the confidentiality of data, selected variables are sufficiently general in character and do not disclose protected information about leasing customers, suppliers and employees, while at the same time they are specific enough to be important for the realization of the model. Figure 1 contains the variables used in the discovery of knowledge from databases. In cases when the sum is smaller than 100%, there were missing data.

Decision trees

Decision trees are a popular and widely accepted tool for classification and prediction, and their strength is reflected in the fact that they are easily understandable due to a graphical display (Apté, Weiss, 1997; Tsang et al., 2011). A decision tree is a statistical method of pattern recognition which is used to solve problems with predictive nature while monitoring the learning process is needed. Predictive problems include forecasting values in the future, pattern recognition, regression of multiple features, the differential analysis, evaluation functions of more features and supervised learning. Decision trees are very efficient when dealing with

large databases and when many variables should be taken into account (Li, 2005; Wu, Banzhaf, 2010).

Figure 1
Variables used in the discovery of knowledge from databases.

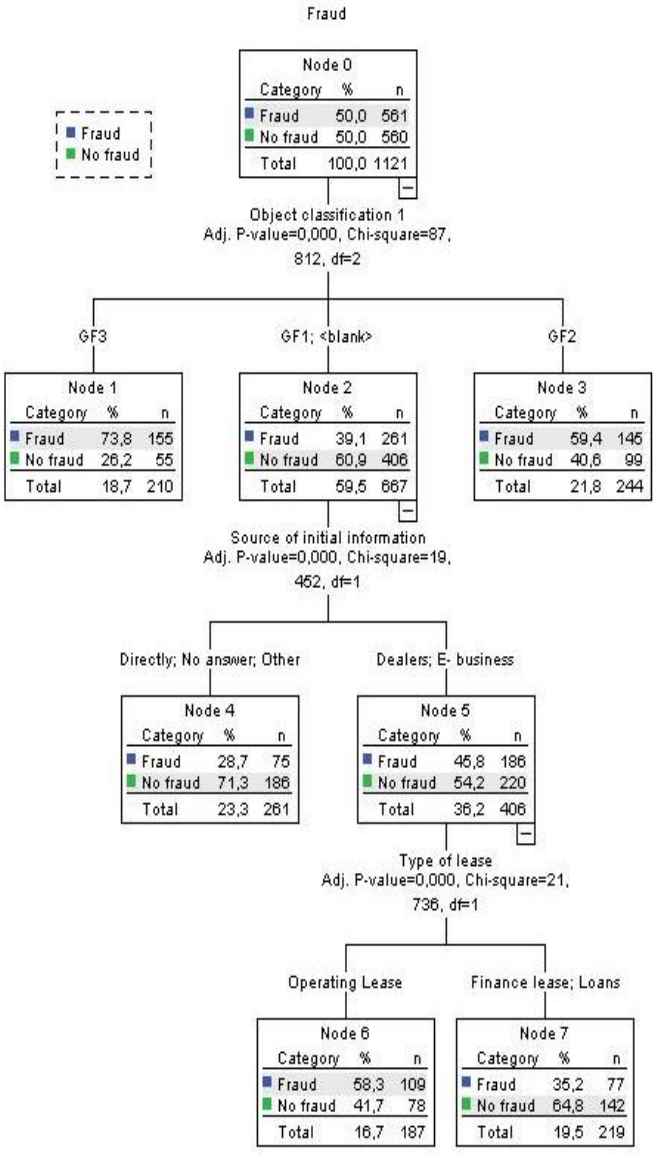


Source: Authors' work

The paper used the CHAID algorithm for trees to detect fraud in the leasing agreements, since this algorithm is suitable for classification problems where the variables have more than two modalities (McCarty, Hastak, 2007; Coussement et al., 2014). The paper uses the software package SPSS, ver. 19th, and two types of models have been developed: (i) Model A: the model with a simpler classification of leased assets (the variable Object classification 1) and (ii) Model B: the model with a complex classification of leasing involving facilities (the variable Object classification 2).

Model A is represented graphically on the Figure 2, and also trough generated business rules in the form of SQL code on the Figure 3.

Figure 2
Decision tree generated with a more aggregate object classification (Object classification 1)



Source: Authors' work

Model A will be described in greater detail. The variable used for branching on the first level is Object 1, which is statistically significant with a level of 1% probability (P-value = 0.000). Second level nodes show branching variables Object 1 at three knots.

- Node 1 (node1) contains 210 data for which the average value of the variable Fraud is 0.738, which means that 73.8% of the agreements for which the subject of the agreement is GF3 resulted in fraud.
- Node 2 has 667 agreements for which the average value of the variable Fraud is 0.391, which means that 39.1% of the agreements for the GF1 and the unknown object contracting resulted in fraud.
- In the same way we interpret Node 3. This node has 244 agreements for which the average value of the variable Fraud is 0.594, which means that the 59.4% of the agreements for the GF2 resulted in fraud.

The variable for branching on the second level is Source of information, which is statistically significant with a probability level of 1% (p-value = 0.000). Third-level nodes show the branching variable Source of information on the two nodes.

- Node 4 shows the clients who come directly to the leasing company or or the source of initial information is not available. This node contains 261 agreements with the average value of 0.287, which means that 28.7% of the agreements resulted in fraud.
- Node 5 shows clients who are contracted through the dealer or the manufacturer, and via the Internet (only a small share). The average value of this node is 0.458, meaning that 45.8% of the agreements resulted in fraud. The variable used for branching on the third level is Type of leasing, which is statistically significant with a probability level of 1% (p-value = 0.000).
- Node 6 contains agreements of operating lease, where the average agreement value is 0.583, meaning that 58.3% of the agreements resulted in fraud. Node 7 includes financial leasing and loans, where the average agreement value is 0.352, meaning that 35.2% of the agreements resulted in fraud.

Figure 3
Rules generated based on decision tree algorithm

```
/* Node 1 */.
IF (Object classification 1 = "GF3")
THEN
Node = 1
Prediction = 'Fraud'
Probability = 0.738095

/* Node 4 */.
IF (Object classification 1 != "GF3" AND Object classification 1 != "GF2") AND (Source of initial information = "Directly" OR Source of initial information = "No answer" OR Source of initial information = "Other")
THEN
Node = 4
Prediction = 'No fraud'
Probability = 0.712644
```

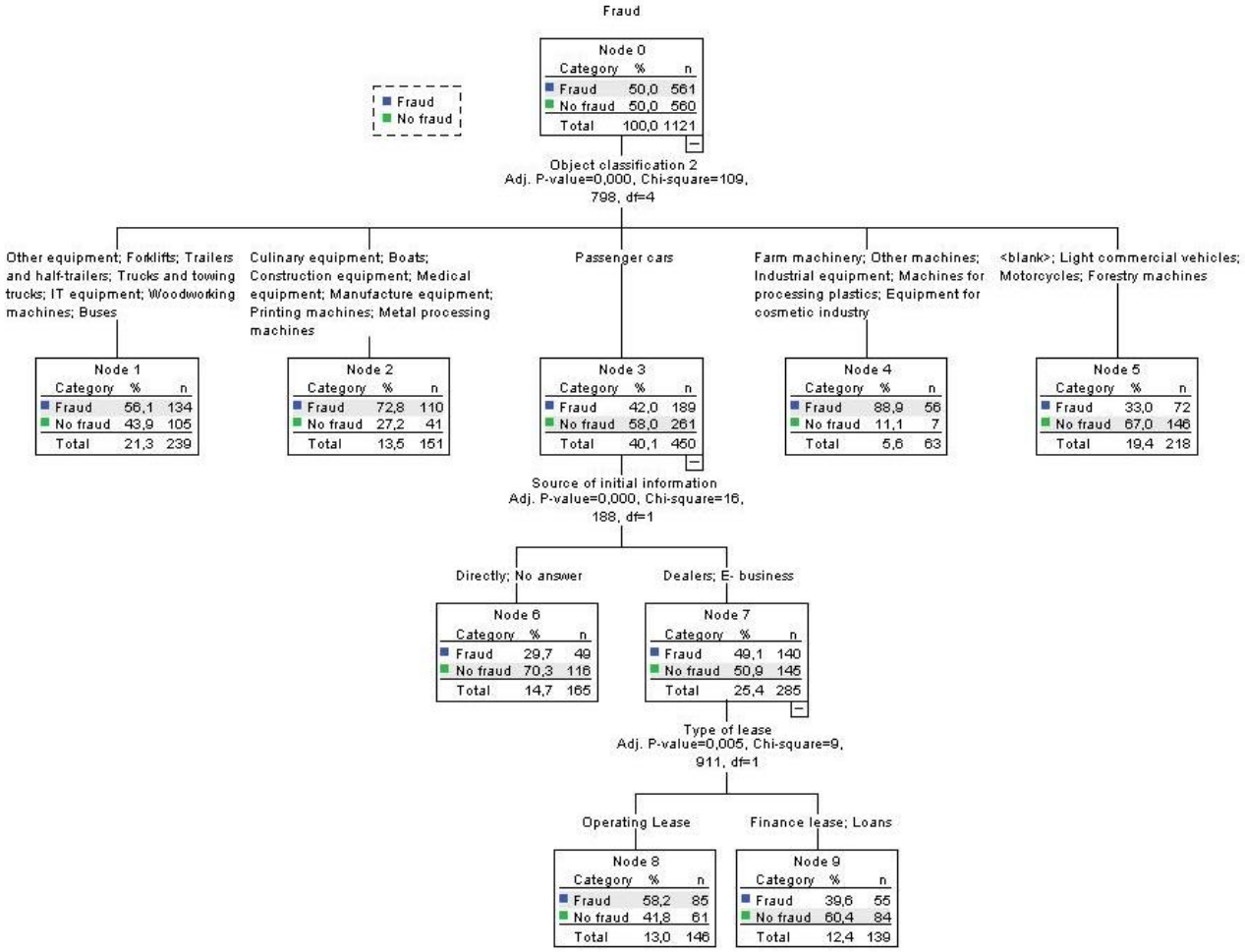
```
/* Node 6 */.
IF (Object classification 1 != "GF3" AND Object classification 1 != "GF2") AND (Source of initial information != "Directly" AND Source of initial information != "No answer" AND Source of initial information != "Other") AND (Type of lease = "Operating Lease")
THEN
Node = 6
Prediction = 'Fraud'
Probability = 0.582888

/* Node 7 */.
IF (Object classification 1 != "GF3" AND Object classification 1 != "GF2") AND (Source of initial information != "Directly" AND Source of initial information != "No answer" AND Source of initial information != "Other") AND (Type of lease != "Operating Lease")
THEN
Node = 7
Prediction = 'No fraud'
Probability = 0.648402

/* Node 3 */.
IF (Object classification 1 = "GF2")
THEN
Node = 3
Prediction = 'Fraud'
Probability = 0.594262
```

Model B is represented graphically on the Figure 4, and also trough generated business rules in the form of rules on the Figure 5. SQL code is provided in the Appendix of the paper.

Model B will be described in greater detail. The variable used for branching on the first level is Object 2, which is statistically significant with a level of 1% probability (P-value = 0.000). Second level nodes are showing branching variables Object 2 at five knots.

- Node 1 (node1) contains 239 data for which the average value of the variable Fraud is 0.561, which means that 56.1% of the agreements for which the subject of the agreement is other equipment, trucks, busses and machines resulted in fraud.
- Node 2 has 151 agreements for which the average value of the variable Fraud is 0.728, which means that 72.8% of the agreements including a wide selection of equipment, machines and boats resulted in fraud.
- Node 3 has 450 agreements for which the average value of the variable Fraud is 0.420, which means that 42.0% of the agreements including passenger cars resulted in fraud.
- Node 4 has 63 agreements for which the average value of the variable Fraud is 0.889, which means that 88.9% of the agreements including farming machines, machines for processing plastics and cosmetic industry resulted in fraud.
- In the same way we interpret Node 5. This node has 218 agreements for which the average value of the variable Fraud is 0.330, which means that 33.0% of the agreements including light commercial vehicles resulted in fraud.

Figure 4
Decision tree generated with a more aggregate object classification (Object classification 1) (Model B)



Source: Authors' work

The variable for branching on the second level is Source of information, which is statistically significant with a probability level of 1% (p-value = 0.000). Third-level nodes show the branching variable Source of information on the two nodes.

- Node 6 shows the clients who come directly to the leasing company or the source of initial information is not available. This node contains 165 agreements with the average value of 0.297, which means that 29.7% of the agreements resulted in fraud.
- Node 7 shows clients who are contracted through the dealer or manufacturer, and via the Internet (only a small share). The average value of this node is 0.491, meaning that 49.1% of the agreements resulted in fraud. The variable used for branching on the third level is Type of leasing, which is statistically significant with a probability level of 1% (p-value = 0.000).
- Node 8 contains 146 agreements of operating lease, where the average agreement value is 0.582, meaning that 58.2% of the agreements resulted in fraud.
- Node 9 includes financial leasing and loan and, contains 139 agreements where the average agreement value is 0.396, meaning that 39.6% of the agreements resulted in fraud.

Figure 5
Rules generated based on decision tree algorithm

/* Node 1 */.
IF ("Other equipment" OR "Forklifts" OR "Trailers and half-trailers" OR "Trucks and towing trucks" OR "IT equipment" OR = "Woodworking machines" OR = "Buses")
THEN
Node = 1
Prediction = 'Fraud' ; Probability = 0.560669

/* Node 2 */.
IF ("Culinary equipment" OR "Boats" OR "Construction equipment" OR "Medical equipment" OR "Manufacture equipment" OR "Printing machines" OR "Metal processing machines")
THEN
Node = 2
Prediction = 'Fraud' ; Probability = 0.728477

/* Node 6 */.
IF ("Other equipment" AND "Culinary equipment" AND "Boats" AND "Forklifts" AND "Trailers and half-trailers" AND "Farm machinery" AND "Light commercial vehicles" AND "Construction equipment" AND "Other machines" AND "Medical equipment" AND "Motorcycles" AND "Industrial equipment" AND "Manufacture equipment" AND "Printing machines" AND "Metal processing machines" AND "Trucks and towing trucks" AND "IT equipment" AND "Machines for processing plastics" AND "Woodworking machines" AND "Equipment for cosmetic industry" AND "Buses" AND "Forestry machines") AND (Source of initial information = "Directly" OR Source of initial information = "No answer")
THEN
Node = 6
Prediction = 'No fraud'; Probability = 0.703030

/* Node 8 */.
IF ("Other equipment" AND "Culinary equipment" AND "Boats" AND "Forklifts" AND "Trailers and half-trailers" AND "Farm machinery" AND "" AND "Light commercial vehicles" AND "Construction equipment" AND "Other machines" AND "Medical equipment" AND "Motorcycles" AND "Industrial equipment" AND "Manufacture equipment" AND "Printing machines" AND "Metal processing machines" AND "Trucks and towing trucks" AND "IT equipment" AND "Machines for processing plastics" AND "Woodworking machines" AND "Equipment for cosmetic industry" AND "Buses" AND "Forestry machines") AND (Source of initial information != "Directly" AND Source of initial information != "No answer") AND (Type of lease != "Finance lease" AND Type of lease != "Loans")
THEN
Node = 8
Prediction = 'Fraud'
Probability = 0.582192

/* Node 9 */.
IF ("Other equipment" AND "Culinary equipment" AND "Boats" AND "Forklifts" AND "Trailers and half-trailers" AND "Farm machinery" AND "" AND "Light commercial vehicles" AND "Construction equipment" AND "Other machines" AND "Medical equipment" AND "Motorcycles" AND "Industrial equipment" AND "Manufacture equipment" AND "Printing machines" AND "Metal processing machines" AND "Trucks and towing trucks" AND "IT equipment" AND "Machines for processing plastics" AND "Woodworking machines" AND "Equipment for cosmetic industry" AND "Buses" AND "Forestry machines") AND (Source of initial information != "Directly" AND Source of initial information != "No answer") AND (Type of lease = "Finance lease" OR Type of lease = "Loans")
THEN
Node = 9
Prediction = 'No fraud'; Probability = 0.604317

```
/* Node 4 */.
IF (Object classification 2 = "Farm machinery" OR Object classification 2 = "Other machines"
OR Object classification 2 = "Industrial equipment" OR Object classification 2 = "Machines for
processing plastics" OR Object classification 2 = "Equipment for cosmetic industry")
THEN
Node = 4
Prediction = 'Fraud'; Probability = 0.888889

/* Node 5 */.
IF (Object classification 2 = "" OR Object classification 2 = "Light commercial vehicles" OR
Object classification 2 = "Motorcycles" OR Object classification 2 = "Forestry machines")
THEN
Node = 5
Prediction = 'No fraud'; Probability = 0.669725
```

Source: Authors' work

Table 2 presents classification matrixes for both Model A and Model B. Surprisingly, Model A is more accurate in predicting fraud, although it uses a more aggregate object classification. Comparison of these models leads to the conclusion that fraud is likely to happen on Object1 - GF3 group, i.e. in the case of Model B – equipment and machinery. This is understandable since these objects of lease have greater value compared to other groups. The logic behind this is that if criminals are going to perpetrate fraud, they will try to maximize the effect. Models also show that firms should be more careful with agreements that come from dealers as there is a higher possibility of fraud. Implementing one of these models or one of their variations would create a good system for fraud detection and could create positive effects on business of a lease company. Implementation of such a solution should be made throughout the industry as a security standard.

Table 5
Classification matrixes for Model A and Model B

Observed	**Predicted**					
	Fraud		No fraud		Percent Correct	
	Model A	Model B	Model A	Model B	Model A	Model B
Fraud	409	385	152	176	72.9%	68.6%
No fraud	232	214	328	346	58.6%	61.8%
Overall Percentage	57.2%	53.4%	42.8%	46.6%	65.7%	65.2%

Source: Authors' work

Practical implications

Introduction of this model in the business would certainly show that certain frauds could be prevented and would indicate the leasing agreements which present a fraud risk. However, to make this project come to life, it would be necessary to develop software that would enable automated, quick and transparent retrieval of data from the system, processing according to the rules and displaying the results in multiple categories. It would be necessary to show already existing fraud events, fraud events that are emerging and potential fraud events so that for each of these categories an appropriate action could be taken.

The solution could be implemented into the current environment through the existing SQL-based applications by developing a separate module. In this case, it would be necessary to employ the original developers to integrate the module within the existing application to set up an alarm system. This is probably the best solution because the program would be incorporated into the existing central application enabling full access to all data in the core system, regardless of the period. According to similar projects, the estimated costs of the development of these modules would be at the level of approximately 15,000 EUR. This estimation is based on the market research conducted for the leasing firm used for the case study. Prevention of even a single case of fraud would prove the purposefulness of this project since instances of fraud in most cases involved expensive leasing objects. Prevention of fraud events results not only in savings connected with the value of lease agreements, but also results in a number of other positive externalities. The accounts receivable department has one less difficult case to handle, there is no need to pay the costs of interventions for finding fraud subjects of leasing and eventually significant legal costs and the costs of hiring legal services staff are avoided.

References

1. Apté, C., Weiss, S. (1997), "Data mining with decision trees and decision rules", Future Generation Computer Systems, Vol. 13, No. 2–3, pp. 197-210.
2. Bhattacharyya, S., et al. (2011), "Data mining for credit card fraud: A comparative study", Decision Support Systems, Vol. 50, No. 3, pp. 602-613.
3. Coussement, K., Van den Bossche, F. A., De Bock, K. W. (2014), "Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees", Journal of Business Research, Vol. 67, No. 1, pp. 2751-2758.
4. Huang, S. Y., Tsaih, R. H., Lin, W. Y. (2012), "Unsupervised neural networks approach for understanding fraudulent financial reporting", Industrial Management & Data Systems, Vol. 112, No. 2, pp. 224-244.
5. Li, X. B. (2005), "A scalable decision tree system and its application in pattern recognition and intrusion detection", Decision Support Systems, Vol. 41, No. 1, pp.112-130.
6. McCarty, J. A., Hastak, M. (2007), "Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression", Journal of Business Research, Vol. 60, No. 6, pp. 656-662.
7. Morais, A. I. (2013), "Why companies choose to lease instead of buy? Insights from academic literature", Academia Revista Latinoamericana de Administración, Vol. 26, No. 3, pp. 432-446.
8. Ngai, E.W.T. et al. (2011), "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature", Decision Support Systems, Vol. 50, No. 3, pp. 559-569.
9. Sinha, A.T., Zhao, H. (2008), "Incorporating domain knowledge into data mining classifiers: An application in indirect lending", Decision Support Systems, Vol. 46, No. 1, pp. 287-299.
10. Smith, C. W., Wakeman, L. M. (1985), "Determinants of corporate leasing activity", Journal of Finance, Vol. 40, No. 3, pp. 895-911.

11. Tsang, S. et al. (2011), "Decision trees for uncertain data", Knowledge and Data Engineering, IEEE Transactions on, Vol. 23, No. 1, pp. 64-78.
12. Wu, S. X., Banzhaf, W. (2010), "The use of computational intelligence in intrusion detection systems: A review", Applied Soft Computing, Vol. 10, No. 1, pp. 1-35.

About the authors

Ivan Horvat received the univ.spec from the Faculty of Economics and Business, University of Zagreb in Information Management. He is currently controlling specialist at the VB leasing Croatia and external associate at Faculty of Economics and Business, University of Zagreb within the area of informatics and SAP. In VB leasing his main focus is on financial controlling, cost control and analysis, budgeting and reporting. Ivan is currently getting additional specialization in internal auditing. Author can be contacted at **ivan.horvat.zg@gmail.com**

Mirjana Pejic-Bach, PhD, is a Full Professor of System Dynamics, Managerial Simulation Games and Data Mining at the Department of Informatics, Faculty of Economics and Business, University of Zagreb. Her current research areas are simulation modelling, data mining and web content research. She is the (co)author of number of articles in international and national journals. She is actively engaged in number of scientific projects (FP7, bilateral cooperation, national projects) and also collaborates in several applied projects in the field of data mining, simulation modelling and informatization. Author can be contacted at **mpejic@efzg.hr**

Marjana Merkač Skok earned her Ph.D. in 1997 from Management and organization sciences at University of Maribor. Currently she is a Dean at Faculty of Business and commercial sciences in Celje, Slovenija. She also works as independent expert for quality assurance in higher education in EU. Before that, she worked as developer and expert in human resource and organizational development in industry and for several years as a business consultant for management. Author is involved in researches about quality, system science, career management, lifelong learning and training. Author can be contacted at **marjana.merkac@fkpv.si**