Bezpieczny Bank
2(59)/2015

Problems and Opinions

Anna Matuszyk[*]
Aneta Ptak-Chmielewska[**]

PROFILE OF THE FRAUDULENT CUSTOMER

1. INTRODUCTION

Fraud may occur in any financial activity. However, banks are particularly exposed due to their role as intermediaries in the financial markets. The risk of financial crime increases concomitantly with an economic downturn, as people are more likely to commit fraud in a recession. This creates significant risk to financial institutions and has recently led to increased interest in proper fraud prevention systems. The key to such systems is to choose the most suitable fraud determinants to identify fraudulent transactions.

Modelling fraud is not the main objective in credit scoring. The main goal is to distinguish good clients from bad ones, without analyzing which of them want to extort money. Over the last decade, there has been growing interest in credit scoring because the number of credit frauds has increased, prompting researchers to look for a solutions to this problem.

According to Dorfleitner and Jahnes (2014), the increasing number of credit defaults caused by application fraud has placed more pressure on banks to maintain the profit of their credit portfolios, since fraud losses are mostly treated as operational risk and result in immediate losses. Furthermore, they are often

[*] Anna Matuszyk is an Assistant Professor at Warsaw School of Economics, Institute of Finance, Warsaw, Poland, Email: anna.matuszyk@sgh.waw.pl
[**] Aneta Ptak-Chmielewska is an Assistant Professor at Warsaw School of Economics, Institute of Statistics and Demography, Warsaw, Poland, Email: aptak@sgh.waw.pl

unexpected and therefore not budgeted, in contrast to classical risk factors based on economic determinants.

In March 2012, the National Fraud Authority published its Annual Fraud Indicator, which estimated that fraud was costing the UK over £73 billion (https://www.gov.uk... 2013). According to CIFAS – the UK's Fraud Prevention Service – motor finance and insurance products each account for roughly 1 in 5 of all application frauds. The Finance Leasing Association (FLA), a trade association for the asset, consumer and motor finance sector in the UK, published figures for motor finance fraud. In the 12 months to September 2011, FLA members reported 840 fraud cases. The value of these cases in terms of the original loan amount was £15.3 million.

In this paper three fraud models were created using the logistic regression, decision tree and neural network approaches. The predictive power of the models was checked using the following measures: percentage of correctly classified cases, ROC curve, Gini coefficient and Average Square Error. The study was based on a real data set consisting of 65,000 personal loans with 350 events of fraud in a bank operating in Europe. The data was provided at the individual level, and the product type was auto loans.

The structure of the paper is as follows. First, we introduce the definition of the fraud event. We outline the main problems encountered when modelling application fraud. In Section 3 we present the available literature in this area. In Section 4 we explain the techniques used in the research, i.e. logistic regression (LR), decision tree (DT) and neural network (NN). In Section 5 we describe the data provided. In Section 6 we explain the details of the models built. Finally, in Section 7 we discuss the results, draw conclusions and outline the possibilities for future research.

2. FRAUD DEFINITION, CLASSIFICATION, PROBLEMS

The definition of a loan application fraud was proposed by Dorfleitner and Jahnes (2014). They distinguished first-, second- and third-party fraud. First-party fraud occurs when a fraudster applies for a loan using his own account and has no intention of repaying the sum. Second-party fraud involves an intermediary who helps to carry out the fraud. And finally, third-party fraud is when a fraudster uses another person's identifying information to perpetrate the crime.

Sandrej (2005) proposed a different classification of fraud, distinguishing internal fraud from external fraud. According to him, external fraud is when the fraudster is outside the bank, while internal fraud is when there is assistance from a bank employee. In a credit card environment there are two main types of fraud: application and behavioural (Bolton, Hand, 2001). When it comes to personal loans, it is application fraud we are dealing with.

There are various reasons why application fraud has not been well researched. One is that it is very difficult to obtain fraud data from financial institutions

because of the need to maintain confidentiality and for competitive reasons. Another reason is the lack of publicly available data. One exception is a small automobile insurance data set used by Phua et al. (2004). There is also a problem with the censorship of detailed results in publications. This is because of the risk that fraudsters could easily use the output to adapt their behaviour.

Another difficulty is related to the data sets, which are usually large, and each transaction must be examined and decisions made in real time. The transactions are often heterogeneous, differing substantially even within an individual account, and the data sets are typically very imbalanced, with only a tiny proportion of transactions belonging to the fraud class (Hand, 2007).

Generally, we can distinguish the following main problems when modelling application fraud:

1) Very limited literature
2) Difficulty in obtaining data
3) Risk of fraudsters changing their behaviour as a result of research findings
4) Fraud data sets are large but only a tiny proportion will be fraudulent transactions.

3. LITERATURE REVIEW

The literature on application fraud in personal loans is very limited. There is some research but mainly into credit card fraud and focusing on behavioural fraud.

A study carried out by Wheeler and Aitken (2000) showed the possibility of using identity information such as names and addresses from credit applications. They used a case-based reasoning approach to analyse the most difficult cases that have been misclassified by existing methods and techniques. An adaptive diagnosis algorithm combining several neighbourhood-based and probabilistic algorithms was found to have the best performance, and the results indicate that an adaptive solution can provide fraud filtering and case ordering functions to reduce the number of required final-line fraud investigations.

A study made by Dorfleitner and Jahnes (2014) was based on a data set consisting of nearly 43,000 personal loan applications from Germany. They found that the sales channel or loan amounts are significant determinants of application fraud. They used a logistic regression method, which was found to be a statistically significant approach for profiling loan application fraudsters. Furthermore, they proved the economic significance of the results by developing a fraud management framework taking into account the fraud rate, the average default cost due to fraud and the costs of fraud screening.

Harmann-Wendels et al. (2009) empirically studied the determinants of new account fraud risk within two dimensions – the probability of fraud, and the

expected and unexpected (monetary) loss-per-account due to fraud. By fraud risk, they mean the risk of a bank failing to enforce a debt because the identity of the person incurring the debt cannot be ascertained. Using a real data set of account applicants, they found that fraud risk is very sensitive to demographic and socio-economic variables such as nationality, gender, marital status, age, occupation and urbanisation. For example, foreigners are 22.25 times more likely to commit account fraud than Germans, and men are 2.5 times more risky than women.

T. Mählmann (2010) studied new account fraud, where an imposter opens lines of credit using a false identity. They analyzed the correlation between fraud and default risk. According to their findings, common socioeconomic/demographic characteristics of account holders have opposite effects on estimated default and fraud probabilities. For example, women possess a lower fraud probability but a higher default probability compared to men and foreigners, who are more likely to engage in account fraud but less likely to default than Germans.

4. METHODS

The following methods were used in creating the fraud models: logistic regression (LR), decision tree (DT) and neural network (NN). Below is a short description each of these techniques.

4.1. Logistic regression

Logistic regression models are a very popular statistical method for predicting customer insolvency. They can be used as binomial models (where one of the variables is dichotomous), or as ordered polynomial ones where the dependent variable can exist in more than two states. Logistic functions can be estimated using the weighted least squares or maximum likelihood method.

The logistic function in the binomial models takes the following form:

$$P(Y = 1) = \frac{1}{1 + \exp^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}},$$

where:
$P(Y=1)$ – dependent variable, in this case it defines the probability of fraud,
β_0 – constant
β_i, $i = 1, 2, \ldots, k$ – weights,
x_i, $i = 1, 2, \ldots, k$ – independent variables.

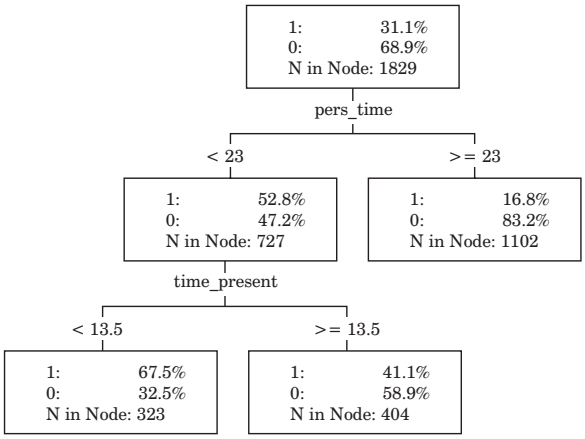
Ratio $P(Y=1)$ takes the values from the interval <0;1>, where 0 is a non-fraudulent customer, and 1 a fraudulent one. The closer to zero value the ratio gets, the lower the probability of

fraud. Logistic regression is a useful tool where the outcome is a binary variable. According to Dorfleitner and Jahnes (2014) logistic regression is a statistically significant approach for profiling loan application fraudsters.

4.2. Decision tree

A decision tree is a non-parametric statistical method. Observations are classified by assigning cases into groups. It calculates the probability of event occurrence at the group level. The decision tree model does not require the prior selection of variables. The main danger when using a decision tree model is the tendency to over-fit, which makes the final model unstable.

Figure 1. Schematic diagram of the decision tree



Source: own elaboration.

The decision tree contains so-called root (the main element, containing the entire data set) nodes and sub-nodes formed by splitting the data according to the rules used. A tree branch creates the node with further subsegments. The final division element is called a leaf, which is the final node and not split further. Each observation of the output file is assigned to one final leaf only. A typical decision tree model, built for a binary dependent variable, contains the following items:

- node definitions – the principles for assigning each observation to a final leaf
- probability (posteriori) for each final leaf which is the ratio of modelled occurrences of the binary variable in each end leaf
- assigned level of the dependent variable in the model for each final leaf.

Decision rules can be based on maximizing profits, minimizing costs or minimizing the misclassification error. In contrast to binary logistic regression,

decision trees do not contain any equations or coefficients, and are based only on the data set allocation rules. The rules generated by the model can be used for prediction without the dependent variable (the result is a binary decision).

After creating a decision tree model with the selected method, the next step is to cut the tree down to the correct size. This is done in stages. Firstly, one division is cut off, then all possible combinations of the trees are checked and the best are chosen. Then another division is cut and the best tree is checked (already shortened twice), etc. As the number of leaves grows, the tree value will initially increase but after reaching a certain point, the growth will not be visible, or a drop can even occur. This is the optimal size of a tree.

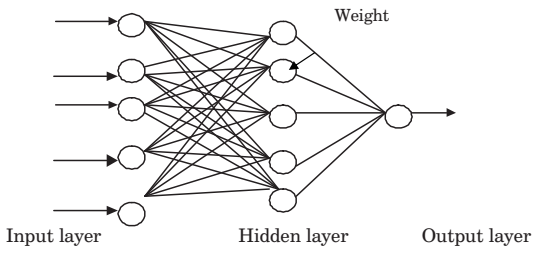
4.3. Neural network

A neural network is one of the methods used in scoring models. In our study, NN should help to specify the relationship between the borrower's characteristics and the probability of fraud. This method also allows you to determine which features are the most important in the fraud event prediction.

A single artificial neuron has multiple inputs x_n, $n=1, 2, \ldots, N$, and one output. Neuron inputs are selected explanatory variables. Indicators are selected based on the method chosen, e.g. the factor analysis method or principal components method. For each variable a specific weight w_n is assigned. Then the total stimulation of the neuron is calculated, which is the sum of the products of the explanatory variables and their weights. The neuron output value depends on the total stimulation of the neuron, which is achieved by using a suitable activation function $\varphi(y)$. The format of this function determines the type of neuron. For a binary variable the activation function for the output layer will be a logistic function, which narrows the estimation to the interval [0:1], making it possible to interpret in terms of the probability of the event occurrence.

The most frequently used is the Multi-layer Perceptron network (MLP network) with one hidden layer (Figure 2).

Figure 2. Schematic diagram of the artificial neural network



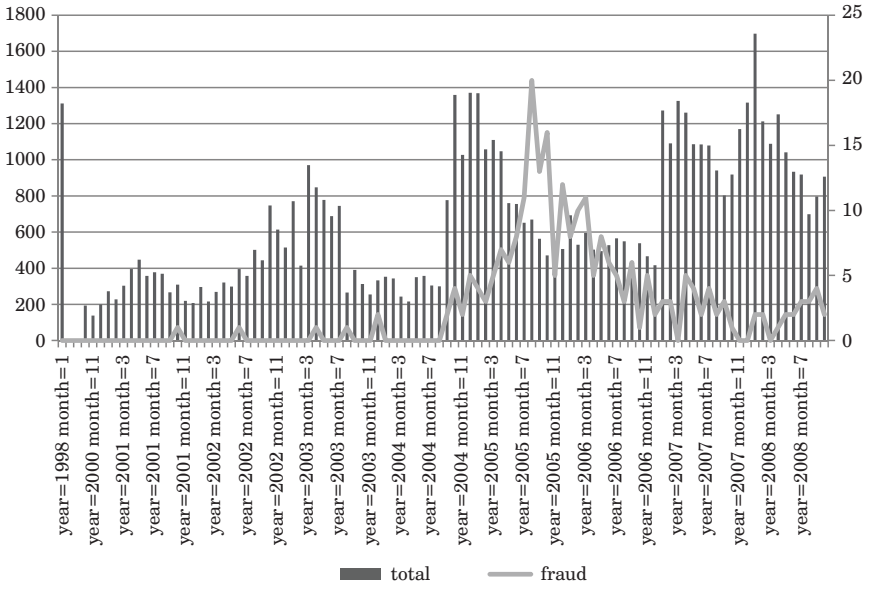
Source: own elaboration.

5. DATA DESCRIPTION

In this study we used a data set from a bank operating in Europe. This dataset covered a period of over 90 months, namely from January 2001 to October 2008. It contains more than 65 thousands cases provided at the individual level. The product type is automobile loans. Due to the small number of fraud events before 2003, all cases before 2003 were deleted. Finally, for modelling purposes, a smaller dataset was used consisting of 980 cases with 245 fraud events. The final sample contains all the fraud cases (245) and 735 randomly selected non-fraud cases, so the proportion is 1:3. This proportion is adequate to measure the first and second type of errors (King, Zeng, 2001).

The fraud definition used by the financial institution that provided the data is as follows: only cases reported to police and courts and then confirmed by the police were considered as fraud events. Figure 3 presents the original data set distribution with the percentage of fraud cases.

Figure 3. Fraudulent transactions in the original data set



Source: own elaboration.

From all the available variables, only those valid at the moment of application were chosen. Table 1 contains a description of the characteristics selected. As a reference category in logistic regression the one with the highest frequency was

selected. All categories with a frequency below 10% of the sample were merged with one another category having a similar fraud rate. Missing data with a frequency lower than 1% was added to the most frequent category.

Table 1. Characteristics used in the models

Characteristic	Description
Brand	SEAT VOLKSWAGEN SKODA (ref. category) OTHER
Category of contract	Annuity (ref. category) Descending/no data
Gender	Female (K) Male (M) (ref. category)
Marital status	he: single/widowed/divorced she: married/widowed she: single/divorced he: married (ref. category)
Commercial phone number given	NO YES (ref. category)
No of scoring	Ordinal: 0,1,2,3,4,5,6
Children	no data/no information no children (ref. category) at least one child
Type of object	USED NEW (ref. category)
Other securities	YES NO (ref. category)
Payment	Direct Debit / no information transfer (ref. category)
Second applicant	YES NO (ref. category)
Type of contract	other standard (ref. category)
Customer	old new (ref. category)
Income Mean £ 0.6 K Median £ 0.5 K	< £ 0.4 K (ref. category) <£ 0.4 – £ 0.7 K) £ 0.7 K +

Characteristic	Description
Financing amount Mean 39,202 PLN Median 33,487 PLN	< £ 5K <£5K–£7K) £7K + (ref. category)
Duration of loan Mean 48.6 months Median 48 months	< 24 months <24–48) months <48–60) months 60 months + (ref. category)
Purchase price Mean £ 10.9 K Median £ 9.4 K	<£7 K (ref. category) <£7 K – £11 K) £11 K+
Downpayment Mean 34 Median 30	< 10% <10–20) % <20–40) % 40%+ (ref. category)
Age	<30 years <30–40) years <40–60) years (ref. category) 60 years +
Year of contract	2003 2004 2005 2006 2007 2008

Source: own elaboration.

Our expectations for the characteristics included are based on the selected sample and refer only to car loans. We expect that customers buying expensive new cars may be susceptible to fraud and may intend not to pay the debt. We would also expect that young people are more risky in comparison to older (retired) customers, so would assume they are high risk. We would also expect that other security measures should make the transaction safer for the bank. Conversely, we would expect older people and families (or at least married customers) to be less risky. The most predictive variable could be the down payment. If the downpayment were high we would expect payments to be made on time. A fraudulent customer would be a new one without any relation to the bank. We would expect the duration of the loan to be a rather neutral variable.

We split the data set into two samples: training and validation. The respective proportions are 75%:25%. Stratified sampling was chosen in order to assure the same proportion of frauds in both samples.

6. RESULTS

In this section we present results obtained from the models built using logistic regression (LR), decision tree (DT) and neural network (NN). Measures were chosen on the basis of those mostly quoted in the literature. All calculations were made using SAS Enterprise Miner and SEMMA methodology.

6.1. Logistic regression

The stepwise selection procedure was applied and variables meeting significance level criteria ($p<0.05$) were chosen to build up the model. Table 2 presents ten final characteristics that were significant in this model.

Table 2. Type 3 effects for logistic regression model

Variable	DF	Chi-sqWald	p-value
Type of contract	1	13.7980	0.0002
Purchase price	2	16.7276	0.0002
Downpayment	3	16.8316	0.0008
Duration of loan	3	12.8616	0.0049
Marital status	3	16.5333	0.0009
Type of object (used/new)	1	15.5664	<.0001
Payment	1	20.8805	<.0001
Second applicant	1	14.8845	0.0001

Source: own elaboration.

According to the results, the significant variables can be divided into three groups:

1) Variables describing the loan type: contract type, method of payment, duration of loan, second applicant, downpayment
2) Variables describing the customer: marital status
3) Variables describing the loan object: type of object, purchase price.

The variable type of contract has two attributes – standard and other. The standard type has 82% lower risk than the other type. As for the method of payment, it can be noticed that direct debit has a lower fraud risk compared to transfer. The length of the loan was another statistically significant predictor in the model. The longer the loan duration, the higher the risk of a fraud event. The largest difference occurs between standard loans (2–4 years) and long loans (over 5

years). The risk in the 2–4 years group is almost 91% lower than in the over 5 years loans group. The next significant variable was the down payment. Loans with an own contribution lower than 10% are 14 times more risky compared to loans with an own contribution over 40%. In the case of the second applicant variable, results obtained were similar to those found by Dorfleitner and Jahnes (2014). A second applicant reduces the fraud risk by almost 86%.

Table 3. Odds ratio for logistic regression model

Variable		Odds ratio	p-value
Type of contract	other standard (ref. category)	0.180	0.0002
Purchase price	£11K + <£7 K – £11K) < £7K (ref. category)	4.500 0.899	0.0061 0.8410
Downpayment	< 10% <10–20) % <20–40) % 40% +(ref. category)	14.114 9.777 3.835	0.0004 0.0005 0.0337
Duration of loan	< 24 months <24–48) months <48–60) months 60 months + (ref. category)	<0.001 0.092 0.255	0.9209 0.0016 0.0539
Marital status	he: single/widowed/divorced she: married/widowed she: single/divorced he: married (ref. category)	5.390 1.008 1.056	0.0006 0.9891 0.9317
Type of object	USED NEW (ref. category)	5.362	<.0001
Payment	Direct debit / no information transfer (ref. category)	0.007	<.0001
Second applicant	YES NO (ref. category)	0.140	0.0001

Source: own elaboration.

Marital status turned out to be a significant variable. The highest risk is from unmarried men. In comparison with married men, the fraud risk in this group is 5.4 times higher. The authors quoted obtained similar results.

Customers buying used cars are over 5 times more risky than customers buying new cars. Dorfleitner and Jahnes (2014) used an additional variable – loan amount – but in our study, purchase price proved to be a much more important variable.

However, the effect on fraud occurrence was similar. The higher the amount, the higher the risk of fraud. Also, the more expensive the car (i.e. costing over £11K), the higher the risk. The risk was 4.5 times higher in compared to the cheaper cars (those less that £7K).

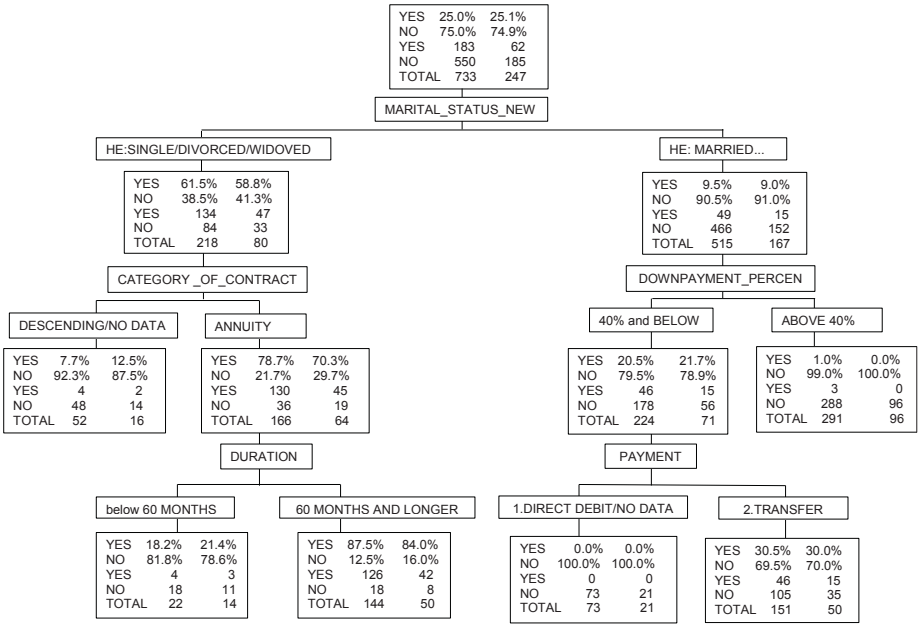
6.2. Decision tree

The significant variables in the decision tree model (assuming significance criteria based on chi-square statistics and significance level 0.2) are as follows in order of priority:

1. Marital status
2. Category of contract
3. Downpayment
4. Payment
5. Duration of loan

The significant variables in this model confirmed the accuracy of the prediction obtained in the regression model. Similar characteristics had a significant effect on the fraud occurrence.

Figure 4. Decision tree path



Source: own elaboration.

Using the result of the decision tree model we were able to define the profile of the typical fraudulent and non-fraudulent customer.

1. *Profile of the fraudulent customer:*
   – man: single / widowed / divorced
   – type of contract: fixed instalments
   – loan duration: 60 + months.

This profile had 150/733 clients (20.4%). The probability assigned to the final leaf in the decision tree model was 86%, which gives a 3.4 times higher risk in comparison to the whole sample (assuming the proportions of frauds in the entire sample equal 25%).

2. *Profile of the non-fraudulent customer:*
   – Woman: married / widow / single / divorced, man: married
   – Downpayment: over 40%.

This profile had 291/733 clients in the training sample (39.7%). The probability assigned to the final leaf in the decision tree model was about 1%, which is almost 25 times lower than in the sample as a whole 1% / 25% = 0.04.

6.3. Neural network (NN)

The results of applying the Neural Network model are presented in Table 4. The Multi-layer Perceptron network was used with one hidden layer and 9 variables included in both the previous models – logistic regression and the decision tree.

Table 4. Results of neural network model

Neural Network Results		
Parameter	**Estimate**	**Gradient Objective Function**
1 CATEGORY_OF_CON1_Descending_noda	–1.076355	–0.000013271
2 TYPE_OF_CONTRACT1_other_H11	–3.343056	0.000043742
3 downpayment_percent1_below10__H1	–0.468720	0.000000653
4 downpayment_percent2_1020__H11	3.492458	–0.000005283
5 downpayment_percent3_2040__H11	–4.604571	0.000016697
6 duration1_24monthsandshorte_H11	–3.528740	0.000010101
7 duration2_2448months_H11	0.988200	–0.000009095
8 duration3_4860months_H11	–0.796016	–0.000016246

Neural Network Results		
Parameter	**Estimate**	**Gradient Objective Function**
9 marital_status_1_he_single_divor	–0.704855	0.000003936
10 marital_status_2_she_married_wid	–2.599170	0.000014920
11 marital_status_3_she_single_divo	0.158492	0.000019830
12 object_used_new1_USED_H11	4.292633	–0.000025460
13 payment1_directdebit_nodata_H11	–2.012679	–5.286869E–8
14 second_applicant1_YES_H11	–0.991820	–0.000029867
15 _DUP	–1.774844	–0.000105000
16 TYPE_OF_CONTRACT1_other_H12	–3.442941	–0.000098794
17 _DUP1	0.244557	–0.000087800
18 downpayment_percent2_1020__H12	1.161893	–0.000115000
19 downpayment_percent3_2040__H12	5.670118	–0.000122000
20 duration1_24monthsandshorte_H12	–1.449947	–0.000110000
21 duration2_2448months_H12	–2.625735	–0.000108000
22 duration3_4860months_H12	–0.130230	–0.000119000
23 _DUP2	2.476361	–0.000114000
24 _DUP3	2.027080	–0.000116000
25 _DUP4	–5.776023	–0.000099894
26 object_used_new1_USED_H12	0.182166	0.000186000
27 payment1_directdebit_nodata_H12	–0.985630	–0.000105000
28 second_applicant1_YES_H12	–4.227913	–0.000087338
29 _DUP5	–0.222298	0.000017854
30 TYPE_OF_CONTRACT1_other_H13	–0.924431	–0.000005365
31 _DUP6	–1.631694	0.000010206
32 downpayment_percent2_1020__H13	1.210802	–0.000003218
33 downpayment_percent3_2040__H13	–0.704159	0.000002633
34 duration1_24monthsandshorte_H13	1.536328	0.000005344
35 duration2_2448months_H13	0.171423	0.000002061
36 duration3_4860months_H13	–1.029980	0.000007026
37 _DUP7	–1.164605	–0.000001681
38 _DUP8	0.647242	0.000008471
39 _DUP9	–0.831810	–0.000012594
40 object_used_new1_USED_H13	0.956127	–0.000005430

Neural Network Results		
Parameter	**Estimate**	**Gradient Objective Function**
41 payment1_directdebit_nodata_H13	1.896231	–0.000025365
42 second_applicant1_YES_H13	0.081030	0.000007703
43 BIAS_H11	–3.921059	0.000037298
44 BIAS_H12	–8.190903	0.000140000
45 BIAS_H13	3.294980	–0.000032590
46 H11_fraudyes	7.803602	–0.000003812
47 H12_fraudyes	2.835943	0.000002616
48 H13_fraudyes	–8.518100	0.000035115
49 BIAS_fraudyes	–1.089161	0.000021293

Source: own elaboration.

6.4. Comparison of the results

All models had similar results (Table 5 and Table 6) but the neural network model was the best one.

Table 5 Comparison of the classification frequencies

Method used	Actual G/ Predicted G	Actual G/ Predicted F	Actual F/ Predicted G	Actual F/ Predicted F
Training sample				
Actual	550	–	–	183
DT	525	25	34	149
LR	526	24	22	161
NN	525	25	13	170
Validation sample				
Actual	185	–	–	62
DT	171	14	10	52
LR	176	9	2	60
NN	173	12	3	59

Legend:
Actual G – actual good customer
Actual F – actual fraudulent customer
Predicted G – predicted good customer
Predicted F – predicted fraudulent customer

Source: own elaboration.

Table 6 presents traditional performance measures, like AUROC, ASE, Gini coefficient and misclassification rate. All the models give very similar results but NN performs best. The misclassification rate for estimated models is very low, at below 10%.

Table 6. Performance measures

Method used	ROC	ASE	Gini Coefficient	Misclassification rate
Training sample				
DT	0.95	0.07	0.90	0.08
LR	0.98	0.05	0.96	0.06
NN	0.99	0.04	0.98	0.05
Validation sample				
DT	0.95	0.08	0.89	0.09
LR	0.98	0.04	0.97	0.05
NN	0.98	0.05	0.96	0.06

Source: own elaboration.

7. CONCLUSIONS

In this study, three models for detecting fraud have been presented. The models were created from real data sets from a financial institution. The model that fits the data best was built on the neural network, however, very low classification errors indicate that the model was overtrained. The logistic regression model was better than the decision tree model (significantly lower classification error for non-fraud events with a similar level of misclassification). In practical usage, the logistic regression model is more beneficial than a neural network or a decision tree model. Nevertheless, the decision tree model provides additional information about the customer profile.

A fraudulent person is most typically a single man (single/divorced/widower) requesting a loan for a five-year period or longer. A detailed screening procedure is definitely not necessary when the customer is a woman (regardless of marital status) or a married man who is applying for an auto loan and has a downpayment greater than 40%.

The conclusions from the models can be used in business practice to reduce costs and save time during creditworthiness analysis. Dorfleitner and Jahnes (2014) described the most risky transactions and tried to give the cut-off point at which it is worth checking the application manually (make a detailed screening) for

transactions that show a significantly high risk of fraud. In our model, we showed the sociodemographic profile of the potentially fraudulent customer which should be of interest during the application procedure. Detailed screening of selected customers makes it unnecessary to use external database screening (in credit bureaus), which gives significant savings. Research will continue in this area using additional data, and new statistical techniques will also be used.

Abstract

When there is an economic downturn, financial crime proliferates and people are more likely to commit fraud. One of the most common frauds is when a loan is secured without any intention of repaying it. Credit crime is a significant risk to financial institutions and has recently led to increased interest in fraud prevention systems. The most important features of such systems are the determinants (warning signals) that allow you to identify potentially fraudulent transactions.

The purpose of this paper is to identify warning signals using the following data mining techniques - logistic regression, decision trees and neural networks. Proper identification of the determinants of a fraudulent transaction can be useful in further analysis, i.e. in the segmentation process or assignment of fraud likelihood. Data obtained in this way allows profiles to be defined for fraudulent and non-fraudulent applicants. Various fraud-scoring models have been created and presented.

Key words: personal loan fraud, fraud determinants, profile of the fraudulent customer

References

Books

Hand, D.J. (2007): *Mining personal banking data to detect fraud.* In *Selected Contributions in Data Analysis and Classification*, ed. P. Brito, P. Bertrand, G. Cucumel, F. de Carvalho, Berlin: Springer, pp. 377–386.

Journals

Bolton, R.J., Hand, D.J. (2002): Statistical Fraud Detection: A Review, *Statistical Sciences* Vol. 17, Issue 3, pp. 235–255.

Delamaire, L., Abdou, H., Pointon, J., (2009): Credit card fraud and detection techniques: A review, *Banks and Bank Systems*, Vol. 4, Issue 2.