Central European Journal of Economic Modelling and Econometrics

Feature Engineering for Anti-Fraud Models Based on Anomaly Detection

Damian Przekop*

Submitted: 18.12.2019, Accepted: 18.06.2020

Abstract

The paper presents two algorithms as a solution to the problem of identifying fraud intentions of a customer. Their purpose is to generate variables that contribute to fraud models' predictive power improvement. In this article, a novel approach to the feature engineering, based on anomaly detection, is presented. As the choice of statistical model used in the research improves predictive capabilities of a solution to some extent, most of the attention should be paid to the choice of proper predictors. The main finding of the research is that model enrichment with additional predictors leads to the further improvement of predictive power and better interpretability of anti-fraud model. The paper is a contribution to the fraud prediction problem but the method presented may generate variable input to every tool equipped with variable-selection algorithm. The cost is the increased complexity of the models obtained. The approach is illustrated on a dataset from one of the European banks.

Keywords: fraud detection, application fraud, feature engineering, anomaly detection, risk modeling

JEL Classification: C550

*Warsaw School of Economics; e-mail: dprzek@sgh.waw.pl; ORCID: 0000-0002-3151-4667

1 Introduction

The phenomenon of fraud is an integral part of human's economic activity. For years fraudulent customers make attempts to outsmart financial institutions in stealing money by applying and not repaying cash loans. Such competition forces banking sector to reduce losses as much as possible. In order to achieve such goal, banks have to predict which customers seem to be fraudulent ones.
Just like in the case of credit risk, bank has to balance the tradeoff between earning money by granting loans and avoiding losses through reducing the volume of loans that are not repaid. However, tools used to control credit risk are unadjusted to detect this part of risk, that is why new additional methodology has to be engaged.
Common approach, both in literature and in practice, is expert investigation of cases most endangered by fraud occurrence. Due to limitations of bank resources, only a part of all applications can be examined, that is why prediction power is important and efficient filtering mechanism has to be used to select only the most risky applications (Maehlmann, 2010). Automatic application rejection can also be considered. Thus anti-fraud analytics is becoming an inevitable part of credit approval process next to well-known credit scoring.
Baesens et al. (2015) enumerate a complete list of components that can be used as a part of analytics in anti-fraud system: supervised and unsupervised modeling techniques, business rules and social networks. Each element targets different types of fraud and combinations of them can be used simultaneously. Supervised models are applied to detect events that happened in the past, unsupervised approach is often focused on anomaly detection and aims finding new patterns of fraudulent behavior, social networks allow for detecting attempts of identity hiding and organized criminal groups and business rules – every remaining fraud pattern.
The rest of this paper is organized as follows. Section 2 reviews related works. In Section 3, research problems are presented. Section 4 and 5 present the proposed approach to the fraud detection problem. Two last sections conclude the paper and provide some future directions.

2 Literature review

The subject of consideration in this paper is application fraud. The literature concerning this topic is very limited. Dorfleitner and Jahnes (2014) explain this state as a consequence of difficulty in obtaining data and censorship of the results obtained. Both factors aim to limit fraudsters' advantage in outsmarting bank security precautions.
Hartmann-Wendels et al. (2009) in their research analyze determinants of application fraud in two dimensions: probability of fraud and loss given fraud. Based on the sample of 200 000 transactions provided by a German internet-only bank they prove dependence of fraud on such demographic and socioeconomic factors as: nationality,

gender, marital status, age, occupation and urbanization. In this research logistic regression is used.
Mählmann (2010) focuses on the cases of first party application fraud. Correlation between credit risk and application fraud risk is examined. He proves that popular socioeconomic and demographic variables have opposite influence on both types of risk. Logistic regression is used.
Dorfleitner and Jahnes (2014) examine factors that have an influence on the probability of committing fraud based on the dataset of 43 000 individual credit applications. Among significant variables are: sales channel and amount of loan. Researchers propose credit fraud management system based on expert investigations of applications that were given high fraud score. Here also logistic regression is used.
Matuszyk and Ptak-Chmielewska (2015) compare fraud detection techniques based on a car loans example. The research involves logistic regression, decision trees and neural networks. The last one offers the best fit.
All papers mentioned above focus on the use of popular demographic and socioeconomic variables and analytical tools widely exploited in the field of credit risk. However there is no discussion on the problem of independent variables' searching and engineering.
In related fields some researchers apply hybrid methods – a combination of supervised and unsupervised approach (Farvaresh and Sepehri, 2011 – telecommunication fraud). In this case outlier detection algorithms are used as a supplement of supervised learning methods through inputting additional predictors based on unsupervised learning experience into supervised methods. Popular methods include: clustering algorithms (Thornton et al., 2014), distance based methods like ODMAD (Koufakou and Georgiopoulos, 2010), density based methods like LOF (Breunig et al., 2000) or INFLO (Jin et al., 2006), entropy based methods (Daneshpazhouh and Sami, 2014).
In some papers manual approach to feature engineering is proposed. Correa Bahnsen et al. (2016) propose fixed transaction aggregates and time features for the problem of credit card fraud detection.

3 Research problems

Fraud risk is a part of operational risk defined in Basel II (2006) as the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events. Application fraud is perceived as a subtype of *external fraud* category proposed in Basel II and is separated from, well described in the literature, credit risk. In spite of the aforementioned definitional split, literature gives no clear distinction between application fraud risk and credit risk in operational sense. Both of them result in the same way – unrepaid loan. Maehlmann (2010) in his paper makes a distinction between both, when analyzing defaulted loans, based on the condition of lacking possibility of identifying the borrower due to identity fraud. According to Dorfleitner and Jahnes (2014), one should make a distinction on three possible

application fraud scenarios: first party fraud, when fraudster commits application fraud using false personal data, second party fraud – additionally involving bank employee and third party fraud – when identity theft occurs. Operationalization of such definition in proposed in Section 4.1.

Having in mind definition of an outlier proposed by Hawkins (1980) being *an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*, and nature of the phenomenon analyzed, application fraud can be perceived as an outlying observation compared to the typical behavior of customers in population analyzed. Here, frauds can be perceived as caused by something different than regular customer's loans. Fraudsters from the very beginning have no intention to repay granted loan. This fact differentiates fraudsters from default customers, i.e. these unable to repay their loan due to effect of credit risk.

In this case, Peer Group Analysis seems to be useful. PGA is defined by Kim and Sohn (2012) as characterizing the expected pattern of behavior around the target sequence by monitoring the behavior of similar objects, and then detecting any differences between the expected pattern and the target. Weston et al. (2008) propose credit card fraud detection algorithm based on tracking of unusual daily behavior of credit card users when compared to their peer group patterns.

Thus a dedicated approach should be formulated in order to improve fraud models' predictive power through utilization of information regarding abnormality of applications. From the bank perspective, such information should be considered in the credit approval process.

Another problem is variable selection. This topic is closely related to rarity of analyzed phenomenon and wide variety of independent variables. Though the dataset consists of 213 variables, in this paper we consider their combinations and transformations. Analyzing all possible interactions involves testing more than $2 * 10^6$ variables. This volume considerably exceeds degrees of freedom at hand, thus information selection should be considered when constructing the final model.

Despite the fact that there are many advanced modeling techniques like neural networks (Dorronsoro et al., 1997 – credit card fraud), Support Vector Machines (Keyan and Tingting, 2011 – money laundering), random forests (Bai et al., 2008 – financial statement fraud) or XGBoost, results of most of them cannot be easily interpreted due to black-box nature of these methods.

Another issue of great importance is proper understanding of model's predictive power improvement. In this case it is model's ability to concentrate possibly many cases of actual frauds among top scores given by model. Such understanding is justified by business application of the analytical solution. Mählmann (2010) suggests that applications with the highest fraud score may be referred for expert investigations performed by analysts who decide whether credit should be granted or not. Due to limitations in human resources that every company has to face, selection of applications most endangered by fraud is inevitable.

4 Experiment

4.1 Data

Dataset originates from European retail bank and consists of only new bank customers' applications (i.e. of these with no product history). 24-month period is used during which 24 107 credit applications were filed. There are 739 frauds that consist of (a) 246 cases of confirmed cash loan application frauds – i.e. where false employment data or false id were identified and (b) 493 based on so called *technical fraud definition*. The idea of *technical fraud definition* is supported by Dorfleitner and Jahnes (2014). In the presented analysis we label as fraud defaulted loans with first nonpayment of an installment taking place before 5% of the loan is repaid. Such customers are perceived as those who had no intention to repay the loan from the very beginning of cooperation. Proposed definition was developed based on literature findings and statistical analysis.
Dataset consists of 213 variables. As the analysis concerns customers with no product history, there is no behavioral data included. The only information accessible when granting a loan is: credit application data and information retrieved from Credit Information Bureau (CIB). Credit application consist of demographic and socio-economic data, household data and employment data. CIB data include credit history of a customer which is a result of his/her contact with the banking sector in Poland. CIB data turns out to be a valuable source of information concerning customer in the sense of number, amount and history of customer's loans – from hire purchase and cash loans to mortgage loans.
Phenomenon concerned is occurrence of fraud on a cash loan. By application fraud we mean confirmed frauds and cases labeled as *technical fraud*. Target variable takes value '1' in 739 cases (3.07% of total).

4.2 Algorithms

The main contribution of this paper is to present two algorithms being a solution to the problem of capturing fraud inclination of a customer. Their purpose is to generate variables that contribute to fraud models' predictive power improvement. These may be represented by a fraudster-specific transformation of features understood twofold: firstly, as a unique combination of features and secondly, as a relative value of a continuous variable within defined peer group.
The first algorithm generates variables that reflect a unique combination of features that is specific to fraudsters. This approach, which utilizes decision trees concept, assumes inputting decision rules as explanatory variables into regression models and can be perceived as a compromise between improvement of algorithm performance via combining results of separate algorithms (like in the case of ensemble methods) and intuitive interpretation of models received. Selected rules are believed to reflect most accurately fraud trends hidden in the data. Logistic regression model is the most

popular approach to the fraud detection problem, however this class of models is based on ceteris paribus assumption and alone cannot capture phenomena dependent on nonlinear transformations of variables. Algorithm proposed in this paper is constructed as follows:

```
p as integer
draw p from (a,b) range

begin
   do set V: draw p variables from the initial data set
      build decision tree T based on set V
   end do

   do rules R:
      define minimal and maximal depth of the rule
      extract rules based on above criterion from tree T
      transform rules into binary variables kept in set R
   end do

   do modeling dataset M:
      examine the dependence between the target and each variable from R
      take variables with p-value below 0.05 in terms of chi-square criterion
   end do
end
```

In the case of presented analysis p is arbitrary set as a random integer from $\{3, 4, 5, 6\}$ set.
The aim of the first algorithm is to find unique combinations of features that may be in correlation with fraud event. Its construction allows to generate variables that are every possible combination of initial dataset variables and can capture more or less obvious predictors that reflect trend frauds. In some cases only combination of features can be predictive in fraud modeling. As an example of features generated by the first algorithm can serve binary variables like: unmarried customer with basic education hired for a short term or customer aged below 20 with master degree. The leaves of the decision tree used should be distanced from the root in the way that the compromise between variable fitting and its support (number of times when it takes value 1) is balanced. That is why the number of levels that constitute depth of the rule should vary from 2 to 4. Only some of the variables proposed can be used as an input to the modeling dataset due to the increasing computation time needed when using more explanatory variables.
The second algorithm is based on the logic of relative value of a continuous variable within defined peer group. This approach is developed based on author's experience in the application fraud detection. It delivers variables indicating unusual features of customers that can be exposed only when compared to the rest of their peer group.

It makes use of relative values of selected features. This algorithm is constructed as follows:

```
n as integer
draw n from (a,b) set

begin
   do set C: draw 1 continuous variable from the initial data set
   end do

   do set D: draw n discrete variables from the initial data set
   end do

   do:
      create disjoint peer groups for every customer based on set D
      do new variables V:
         first one: a ratio of continuous variable value to its
                    mean value within the peer group,
         second one: a value of a continuous variable's distribution
                     function within the peer group
      end do
   end do

   do modeling dataset M:
      examine the dependence between the target and each variable from V
      take variables with p-value below 0.05 in terms of chi-square criterion
   end do
end
```

In the case of presented analysis p is arbitrary set as a random integer from $\{1, 2, 3, 4\}$ set.
The second algorithm targets information helpful in fraud detection that can be noticed only within the most comparable peer group. The same absolute level of variable (for example salary) can mean two different things in two different peer groups. The aim of the algorithm is to find out the relative meaning of continuous variables' values.
Variables generated using both algorithms are then used in the modeling part of the analysis. Their purpose is to represent untypical features of fraudulent customers when compared to the rest of the population. The research hypothesis is that setting variables generated by both algorithms as a part of modeling dataset contributes to the improvement of fraud model's predictive power through utilization of unique information regarding fraud trends that cannot be tracked in one-dimensional analyses. New features allow a better understanding of the phenomenon and improve interpretability of anti-fraud models.

4.3 Experiment

The first step of an experiment involved logistic regression supplementation with outlier score derived based on methods mentioned in Section 2, namely: LOF, INFLO, ODMAD and clustering approach. Additional predictors were supposed to improve anti-fraud models predictive power in terms of Lift 5% measure. Consequently, 60 new predictors based on LOF algorithm were added (derived based on different algorithm parametrization), 60 new predictors based on LOF algorithm (just like in the case of LOF), 5 new predictors based on ODMAD algorithm (parametrizations that generated scores with lowest ex post rates of false positives) and 14 based on k-means clustering algorithm (7 runs for divisions on 4-10 clusters and 7 measures of distance to clusters' centroids).
None of the predictors mentioned above contributed to the improvement of model's predictive power in terms of Lift 5% measure. That is why contribution of proposed algorithms is verified. Variables generated by both of them are used as additional input to the modeling dataset.
In order to test the usefulness of the proposed approach to improvement of fraud models' predictive power described above, LASSO logistic regression (least absolute shrinkage and selection operator) described by Tibshirani (1996) is used. Algorithm proposed above is of general purpose and other modeling techniques can be used instead. Thus model selection is not of major importance and key role, in the conducted research, belongs to the construction of variables described in the previous section.
The main goal of logistic LASSO regression is to solve the following optimization problem (1):

$$\min_{\theta} \sum_{i=1}^{N} -\log p\left(y^{(i)} \mid x^{(i)}; \theta\right), \quad p.w. \sum_{j=1}^{M} |\theta_j| < C, \tag{1}$$

where N observations $\{\left(x^{(i)}, y^{(i)}\right), i = 1, ..., N\}$ are described with M-dimensional vector of explanatory variables $x^{(i)} \in \mathbb{R}^M$ and $y^{(i)} \in \{0, 1\}$. For low values of t, some of parameters computed by the algorithm take value 0. This feature of the algorithm is of major importance especially due to the fact that number of variables available exceed number of observations.
Modeling approach described above is applied respectively: to dataset consisting of only initial variables (A) and datasets supplemented with variables proposed by the described algorithms (B-G). In the analysis following datasets are used:

- Set A – consisting only of initial variables,
- Set B – consisting of initial variables and variables generated by 2^{nd} algorithm (based on distribution of continuous variables within peer groups),
- Set C – consisting of initial variables and variables generated by 2^{nd} algorithm (based on mean values of continuous variables within peer groups),

- Set E – being sum of sets: B and C,

- Set F – consisting of initial variables and variables generated by 1^{st} algorithm (based on the concept of decision tree rules utilization),

- Set G – being sum of sets E and F.

Sets B, C and F are supplemented with 5 000 variables generated by both algorithms that are in strong relation with target variable in terms of chi-square criterion. The decision on number of variables added is arbitrary and in this case is determined by computation time needed when more explanatory variables are used.
As a measure of model's predictive power improvement Lift 5% measure (the value of Lift for 5% of top scored observations) is chosen. This measure reflects how well the model fulfil its business requirements in terms of prediction. The cases given highest fraud score can be referred for expert investigation performed analysts who decide on loan granting or even automatically rejected.
Consequently there is a need to select such group of applications that is not only most endangered with fraud risk but also is characterized by high concentration of applications filed by fraudsters. Imprecise automatic rejection of applications can result in bank's revenue reduction and in case of expert investigation – huge personal costs that bank has to pay.
In modeling part undersampling 10:90 is used. This means that modeling sample is based on all fraud events and nine times more drawn non-fraud cases. The concept of undersampling is recommended in fraud modeling due to rarity of this phenomenon (Baesens 2015). The sample is divided into training and test set using 60:40 ratio. Test set is used to assess stability of the proposed models. Verification of solution's stability is a vital part of the analysis due to the risk of model's overtraining what leads to drop in model's ability to generalize predictions.

5 Results

For all the sets enumerated above LASSO logistic regression was applied. Lift 5% measure for such models is presented in Table 1.

Table 1: Values of Lift 5% measure for models A-G

	A	B	C	E	F	G
training set	3.34	4.65	4.11	4.88	6.19	5.96
test set	3.58	3.58	4.19	4.32	4.53	4.46

According to Table 1, model F has the highest value of Lift 5% measure. In spite of the fact that model G is better adjusted to the training set (regression was run on dataset with more variables), it has inferior predictive power than solutions based

on smaller group of variables. It should be noted that all the models proposed in the experiment deliver better predictions on the test set than the model based on initial variables does on training set. This means that new variables contribute to improvement of fraud models' predictive power.
Taking into consideration differences on values of Lift 5% levels between training and test sets across all presented models, some overfitting can be observed on supplemented models.
Lift distribution for training set is presented in Figure 1 and for test set in Figure 2.

Figure 1: LIFT curves for models A-G (training set)

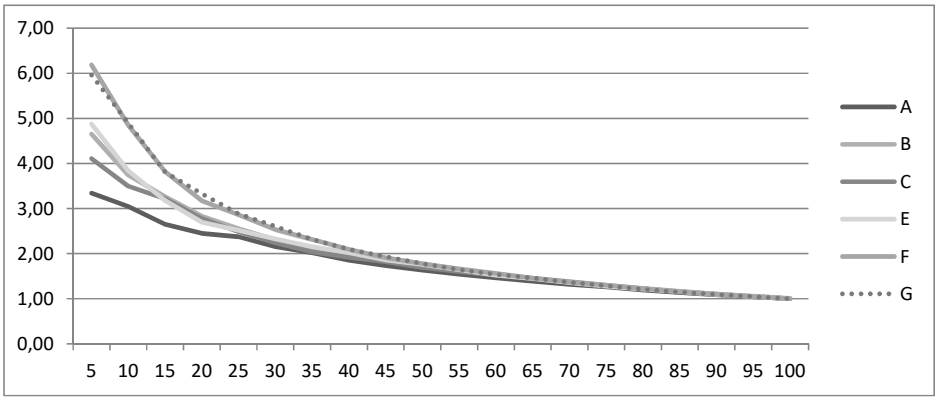


Figure 2: LIFT curves for models A-G (test set)

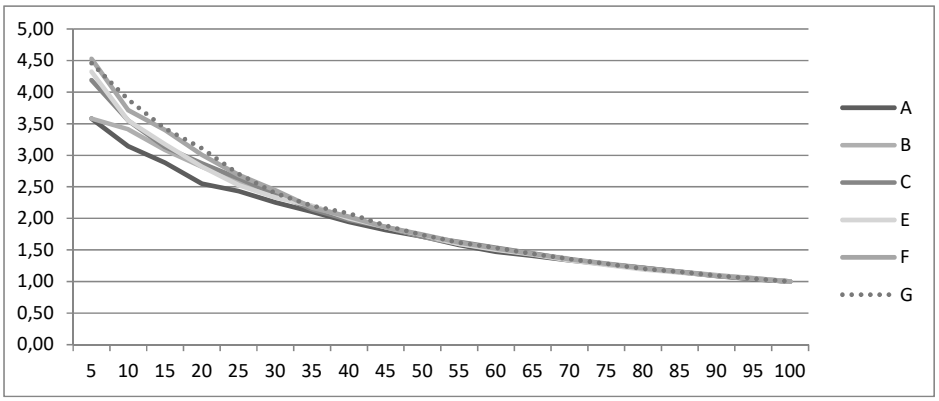


Figure 1 shows that for training set models F and G deliver improvement in prediction power respectively by: 85% and 78% in terms of the value of Lift 5% measure. Figure 2, presenting the models' ability of result generalization shows that models F and G deliver improvement in predictions respectively by: 26% and 25%.

5.1 Sensitivity analysis

Due to the risk of models' overfitting, solution's sensitivity to the final choice of variables is examined. Table 2 contains numbers of variables included in each of the presented models.

Table 2: Numbers of variables included in models A-G

A	B	C	E	F	G
14	86	34	65	98	119

The content of Table 2 shows that the improvement of predictive power between models: G and A is followed by eight times increase in the number of variables used. In order to verify the influence of variables reduction on quality of prediction backward selection regression is applied. Significance level staying criterion is set to $\alpha = 0.01$. As a result values of Lift 5% measure for models obtained are presented in Table 3 and numbers of variables used are presented in Table 4.

Table 3: Values of Lift 5% measure for models A-G with reduced number of variables

	A	B	C	E	F	G
training set	3.75	4.88	4.42	4.88	6.46	6.82
test set	3.92	3.92	4.19	3.99	4.59	4.66

Table 4: Numbers of variables included in models A-G with reduced number of variables

A	B	C	E	F	G
10	25	16	19	40	49

According to the content of Tables: 3 and 4, model G, after 2.5 times reduction in number of variables, presents the best predictive properties in terms of Lift 5% measure value on test dataset. It should be noted that reduction in number of variables does not impact negatively models' predictive power.
Lift distribution of the reduced models for training set is presented in Figure 3 and for test set in Figure 4.

Figure 3: LIFT curves for models A-G with reduced number of variables (training set)

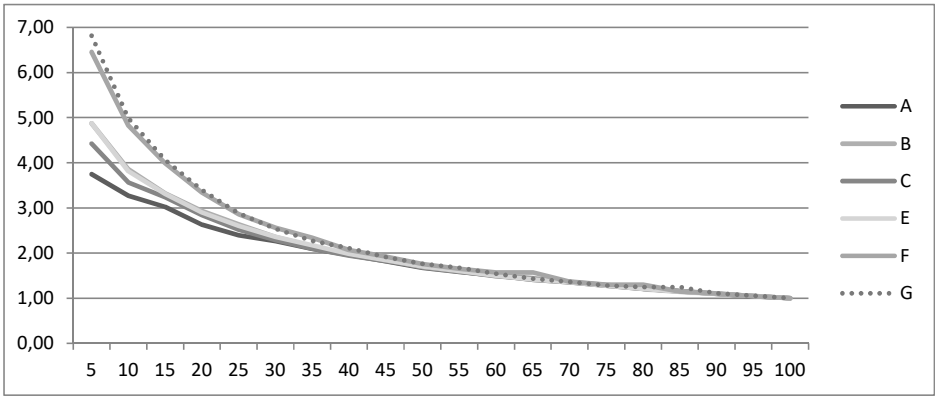
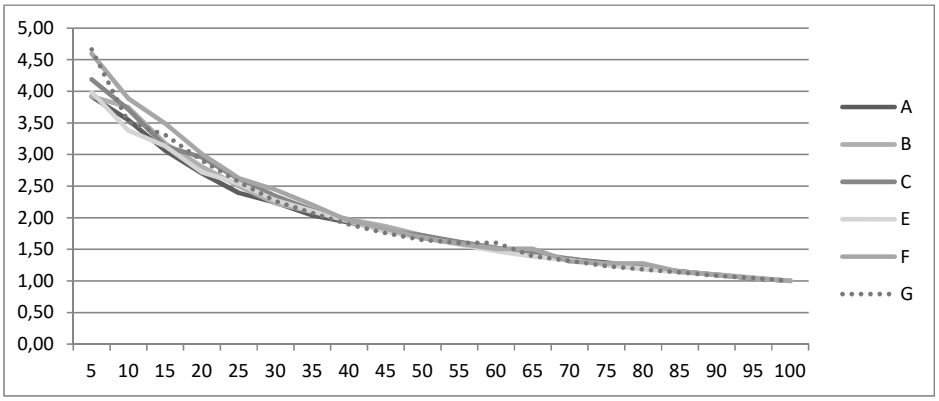


Figure 4: LIFT curves for models A-G with reduced number of variables (test set)



As shown in Figure 3, models F and G deliver improvement in prediction for training set respectively by: 72% and 82% in terms of Lift 5% measure value. Figure 4, presenting the models' ability of result generalization shows that models F and G deliver improvement in prediction respectively by: 17% and 19%.
Models that are significantly outperforming in terms of Lift 5% measure value are these based on sets: F and G. What these sets have in common is input consisting of variables generated by algorithm based on the concept of decision tree rules utilization. These regressors seem to have the strongest influence on fraud models' predictive power improvement.
Synergy between both algorithms can be observed. Lift 5% for model based on E dataset (initial data variables from the second algorithm) is 3.99, for model based on

F dataset (initial data variables from the first algorithm) is 4.59. However, Lift 5% for model based on joint G dataset is at higher level of 4.66.

5.2 Stability analysis

In order to verify influence of sample selection on the results presented above, stability analysis was performed. It assumed 1000 simulations based on randomly selected samples. Each time 6 models based on A-G sets were built on 95% of frauds and on good loans in 10:90 proportion. For each model Lift 5% measure was calculated on training and test sets. Results were averaged and put in Table 5.

Table 5: Values of Lift 5% measure for models A-G (stability analysis)

	A	B	C	E	F	G
training set	4.25	4.29	4.26	4.42	5.97	5.98
test set	3.86	3.80	3.77	3.76	4.74	4.81

According to Table 5, mean values of Lift 5% measures are similar for A-E models on test sets. Significantly higher values are attributed to F and G models. What these sets have in common is input consisting of variables generated by algorithm based on the concept of decision tree rules utilization. These regressors seem to have the strongest influence on fraud models' predictive power improvement. Here the effect of synergy between both algorithms can be observed too. Models F and G deliver improvement in prediction respectively by: 23% and 25% in terms of Lift 5% measure.

5.3 Interpretations

Overview of variables that are included in model G can be a source of information regarding influence of the solution proposed in this paper on models' predictive power improvement. Among 49 variables that make up model G, 3 come from initial dataset, 43 are generated by algorithm based on the concept of decision tree rules utilization, 3 are generated by algorithm based on the concept of continuous variable's relative value within defined peer group. Exemplary variables of model G are:

- ratio of applicant's age to number of loans applied during the last 3 months within the peer group being a combination of educational stage and fact of delay in repaying the obligations during the last 12 months,

- the fact that customer applies for a loan exceeding 10 000 EUR and his/her marital status is single,

- the fact that customer is aged < 20 and his/her marital status is not single.

Coefficients of the second and the last variable are positive whereas of remaining ones – negative. In spite of the complicated construction, these variables seem to be logical and have business interpretation.
The first one indicates the self-responsibility of a customer – number of loans taken by a customer may increase with age, however young customers with many loans are endangered by fraud risk. Such ratio should be benchmarked against the most comparable group of clients – in this case it is determined by educational stage and punctuality in installments paying. The second one shows that people with relatively less obligations (single marital status) may be more prone to take the risk of stealing huge amounts of money. The last one represents young people that have more obligations than their single colleagues and are less prone to risk well-being of their families by stealing bank's money.
Significance of the new features in model G proves that the proposed algorithms generate variables that capture specific relations between probability of committing fraud and predictors at hand. Presented way od feature engineering covers nonlinearities and interactions that influence the target variable simultaneously. Advantage of this approach is interpretability of these effects.

6 Summary

Motivation to conduct this research was to verify whether predictive power of fraud model can be improved through adding new variables, formulated based on proposed algorithms.
In this paper, we proved that outlier detection techniques are too general and unadjusted to the problem of detecting application frauds. It turns out that application fraud is so specific phenomenon that requires dedicated approach. Finding abnormality of fraudulent applications requires searching of specific transformation of features that can be used in solving predictive problem.
Two algorithms generating unique predictors were proposed. It was proved that their application contributes to substantial improvement of fraud models' predictive power. Solution described in this paper is of general use in the sense that it may generate variable input to every tool equipped with variable-selection algorithm.
Despite the fact that there are some modeling techniques that offer the same or even better results than the proposed solution in terms of predictive power, new features allow a better understanding of the phenomenon and improve interpretability of anti-fraud models.
According to the results, expected improvement of fraud models' predictive power was obtained. Based on Lift 5% measure, the prediction gets better by 25% on the test sample.
In spite of the models' overtraining, it can be offset by the reduction of the volume of variables used. Even after that the level of predictive power improvement remains satisfactory.

7 Discussion

Approach presented above requires additional analyses. As far as future work is concerned, time stability and sample selection analyses should be conducted and a deeper insight into the trade-off between model quality improvement and reduction of number of variables used should be made. In addition, it would be interesting to verify performance of the proposed algorithms for other rare-event problems than fraud modeling.
Another interesting direction for future work is combination of both above-mentioned algorithms into one integrated solution. According to preliminary results obtained by author, such solution offers better prediction results. However, it involves also further computational complexity.

References

[1] Baesens B., Van Vlasselaer V., Verbeke W., (2015), *Fraud Analytics using descriptive, predictive and social network techniques*, Wiley and SAS Business Series, 1st Edition.

[2] Bai B., Yen J., Yang X., (2008), False financial statements: characteristics of China's listed companies and CART detecting approach, *International Journal of Information Technology & Decision Making* 7, 339–359.

[3] Basel Committee on Banking Supervision, (2006), *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*, Bank for International Settlements, Basel.

[4] Breunig M. M., Kriegel H. P., Ng R. T., Sander J., (2000), *LOF: identifying density-based local outliers*, ACM Sigmod Record.

[5] Correa Bahnsen A., Aouada D., Stojanovic A., Ottersten B., (2016), Feature Engineering Strategies for Credit Card Fraud Detection, *Expert Systems with Applications* 51.

[6] Daneshpazhouh A., Sami A., (2014), Entropy-based outlier detection using semi-supervised approach with few positive examples, *Pattern Recognition Letters* 49, 77–84.

[7] Dorronsoro J., Ginel F., Sanchez C., Cruz C., (1997), Neural fraud detection in credit card operations, *Neural Networks* 8, 827–834.

[8] Dorfleitner G., Jahnes H., (2014), What factors drive personal loan fraud? Evidence from Germany, *Review of Managerial Science* 8(1), 89–119.

[9] Farvaresh H., Sepehri M., (2011), A data mining framework for detecting subscription fraud in telecommunication, *Engineering Applications of Artificial Intelligence* 24(1), 182–194.

[10] Hartmann-Wendels T., Mählmann T., Versen T., (2009), Determinants of banks' risk exposure to new account fraud – Evidence from Germany, *Journal of Banking & Finance* 33, 347–357.

[11] Hawkins D., (1980), *Identification of Outliers*, Chapman and Hall Hawkins, London.

[12] Jin Y., Rejesus R. M., Little B. B., (2005), Binary choice models for rare events data: a crop insurance fraud application, *Applied Economics* 37, 841–848.

[13] Keyan L., Tingting Y., (2011), An Improved Support-Vector Network Model for Anti-Money Laundering, *Fifth International Conference on Management of e-Commerce and e-Government (ICMeCG).*

[14] Kim Y., Sohn S., (2012), Stock fraud detection using peer group analysis, *Expert Systems with Applications* 39, 8986–8992.

[15] Koufakou A., Georgiopoulos M., (2010), A fast outlier detection strategy for distributed high-dimensional datasets with mixed attributes, *Principles of Data Mining and Knowledge Discovery* 20, 259–289.

[16] Mählmann T., (2010), On the correlation between fraud and default risk, *Zeitschrift für Betriebswirtschaft*, December, 80(12), 1325–1352.

[17] Tibshirani T., (1996), Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1), 267–288.

[18] Weston D., Hand D., Adams N., Whitrow C., Juszczak P., (2008), Plastic card fraud detection using peer group analysis, *Advances in Data Analysis and Classification* 2(1), 45–62.