QUANTITATIVE METHODS IN ECONOMICS
Volume XXV, No. 2, 2024, pp. 58 – 69

EFFECTIVENESS OF VARIABLE SELECTION METHODS FOR MACHINE LEARNING AND CLASSICAL STATISTICAL MODELS

Urszula Grzybowska https://orcid.org/0000-0001-7342-5382
Marek Karwański https://orcid.org/0000-0001-5192-7920
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
Instytut Informatyki Technicznej
e-mail: urszula_grzybowska@sggw.edu.pl; marek_karwanski@sggw.edu.pl

Abstract: In line with new international financial supervision directives (IFRS9), banks should look at a new set of analytical tools, such as machine learning. The introduction of these methods into banking practice requires reformulation of business goals, both in terms of the accuracy of predictions and the definition of risk factors. The article compares methods for selecting variables and assigning "importance" in statistical and algorithmic models. The calculations were carried out using the example of financial data classification for loan default. The effectiveness of various machine learning algorithms on selected sets of variables was compared. The results of the analyzes indicate the need to revise the concept of the "importance" of a variable so that it does not depend on the structure of the model.

Keywords: variable selection, machine learning, variable importance

JEL classification: C45, C52, C55

INTRODUCTION

Classical statistical methods have had well established model selection measures and variable significance tests for decades now. Model selection can be done based on AIC or BIC criteria. On other hand variable significance can be obtained in machine learning models.

Machine learning models are better suited for large data sets with many observations.

https://doi.org/10.22630/MIBE.2024.25.2.6

In this paper we present the application of different methods for variable selection using a loan defaults dataset with 88 variables as an example. Our aim is to present and apply contemporary variable selection methods in an economic context and compare it with classical statistical approach. We show the application of methods such as LASSO, Ridge or recursive feature elimination, and the determination of variable importance in algorithmic models: Random Forests, Gradient Boosting, XGBooost and Neural Networks (NN). We also investigate the impact of the number of observations on the variable selection process. For the resulting sets of variables, we compare the effectiveness of the algorithmic methods and logistic regression. In the first section of the paper, we present different ways of selecting variables for the model, together with a review of the literature on this issue. In the second section, we describe the data. In the third chapter, we present computational results on both the selection of variables and the performance of the machine learning models on the considered sets of variables. The last chapter is devoted to conclusions and a summary.

The calculations were performed in Python ver. 3.9.

METHODS

Variable selection related to data dimension reduction is crucial. It allows the elimination of irrelevant, redundant variables, avoids overtraining the model, increases computational speed and allows better interpretation of results. There are many strategies of variable selection and an overview of them can be found in numerous publications (see [Bag et al. 2022, Li et al. 2017; Pudjihartono et al. 2022; Jia et al. 2022; Zebari et al. 2020, Sauerbrei et al. 2020]). The methods can be divided into methods related to the data model so-called wrapper methods (e.g. recursive feature elimination, heuristic methods) or embedded methods (Random Forests, LASSO, Ridge Regression) [Lal et al. 2006] and model-independent filtering methods [Sánchez-Maroño et al. 2007; Hopf 2021] e.g. based on variable correlation or mutual information (MI) measures [Vergara 2014; Gajowniczek et al. 2022].

Statistical modelling

Hypothesis testing is the most common criterion for variable selection in practical statistical modelling problems. Iterative testing of models is performed through forward selection or backward selection algorithms, depending on whether one starts with an empty model or a model with all variables that can be considered. While significance criteria are typically used to include or exclude variables from a model, information criteria focus on selecting a model from a set of plausible models. Including more variables in the model increases the fit of the model. Unfortunately, such a fit is not always desirable, as it leads to an increase in fitting error. Information criteria have been developed to avoid this apparent fitting effect leading to the selection of more complex models. For example, AIC or BIC statistics are used as information criteria.

Variable selection can also be carried out using a strategy based on the so-called regularisation operator (LASSO), which involves imposing additional conditions on the error function when calculating regression coefficients [Hastie et al. 2008; Hastie et al. 2015, p. 32]. LASSO models are widely used in multivariate problems. Regression coefficients estimated by LASSO procedures are biased but may have a smaller mean square total error than by conventional estimation. Because of the loading, their interpretation in explanatory or descriptive models is difficult, and confidence intervals based on resampling procedures such as bootstrap do not reach their stated nominal level [Taylor, Tibshirani 2015].

Variable selection in algorithmic modelling

In the case of algorithmic models, dedicated ways of assigning importance to predictor variables are created. [Elith et al 2008; Adler, Painsky 2022]. Values that measure the weights of variables help in the interpretation of the data, as well as ranking the variables and facilitating the selection of variables into the model. Within machine learning models such as Decision Trees, Random Forests, Gradient Boosting, XGBoost or LightGBM variable importance is measured based on the number of times a variable is selected for a split [Elith et al 2008; Ben Jabeur et al 2023]. CatBoost [Dorogush et al. 2017; Ostroumova et al. 2018] is a relatively new tree based algorithm that has been designed to deal with categorical features. Unlike other tree based algorithms, it does not require categorical features to be one-hot-encoded. As a result it enables easier interpretation of feature importence. Within algorithmic modelling, we can also use one of the wrapper methods, recursive feature elimination (RFE). In this method, smaller and smaller subsets of variables are considered, the least important variables are removed from the current set based on a measure of importance, until a predetermined number of variables is reached [Kohavi, John 1997; Priyatno et al. 2024]. The approach can however be computationally expensive. For neural networks, one way to select variables is the VIANN method based on a modification of Welford's algorithm [De Sa 2019]. To assess the validity of the variable x_s, we use a measure based on the averaged variance of the changes in the weights-parameters of the first hidden layer network connected to the input variable x_s throughout the back-propagation process. This means that the final validity score of the variable will depend on both the final weights-parameters and the variance of them during training. It is assumed that the more the weight $w_{(a,b)}$ of the connection (a, b) varies during the learning phase, the greater the importance of node a in the prediction process. Using VIANN, we need to determine at which learning stages we update the variance. Several options can be considered, due to iteration (after each batch), per epoch or user-defined interval. For simplicity in the paper, we update the variance of the weights at each epoch. Feature importance ranking (FIR) for deep learning models has been described in [Wojtas, Chen 2020].

In addition to the way variables are selected within machine learning methods, there are also hybrid methods that, for example, combine filtering methods with methods based on machine learning approach. Recently, heuristic methods for selecting variables for the model have been used for large sets [Jia et al. 2022].

We will use a hybrid model in our work. We will first eliminate highly correlated variables, quasi-constant variables and variables with high VIF. For the remaining continuous variables, we will select variables using the LASSO method and Ridge Regression in Logistic Regression. Logistic Regression enables regularization that helps to avoid overfitting and can be used for variable selection similarly as it is in linear regression. We also calculate the importance of variables in common machine learning models.

DATA DESCRIPTION

In the research we have used a loan defaults dataset. The set contained 155 572 observations, among them 15802 defaulted. The observations were described by 88 continuous variables. Among the variables 15 were created artificially and had no impact on the target variable Default. The following subsets were used for the calculations:

- Set 1: 250 observations
- Set 2: 500 observations
- Set 3: 1000 observations.

For each set seperately relevant, i.e., influencial features were searched for.

RESULTS

There were 88 continuous variables in the database. In a first step, correlated variables were eliminated, taking a threshold value of 0.9. Variables with a high VIF were also removed, taking a threshold value of 20. Also quasi-constant variables were removed. The remaining variables were taken for further analysis. Further variable selection was done using machine learning methods and also using L1 regularisation (LASSO) and L2 regularisation (Ridge regression, Ridge) in logistic regression, which are embedded methods. The importance of variables was determined by machine learning methods, i.e. using Random Forest (RF), XGBoost algorithm (XGB), Gradient Boosting (GB) and Neural Networks (NN). As a result of a preliminary elimination of correlated variables, varaiables with low variability (quasi-constants), and variables with high VIF, we have obtained the following sets of features: for Set 1 there were 57 explanatory variables distinguished, for Set 2 there were 58 explanatory variables left and for Set 3 there were 49 explanatory variables left.

Statistical feature extraction

A logistic regression model was built for each set of variables in turn, taking into account the Firth correction [Firth 1993; Puhr et al. 2017] for sets 2 and 3. Forward selection for Set 1 performed for 57 distinguished variables left hardly 3 significant variables. The results are shown in Table 1. The AUC for this model is 1 which indicates perfect classification.

Table 1. Results of logistic regression performed for Set 1

Summary of Forward Selection					
Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	x38	1	1	155.701	<0.0001
2	x39	1	2	87.8676	<0.0001
3	x34	1	3	6.2232	0.0126

Source: own calculations

Model built for x38 and x39 with Firth correction provided the following results with AUC=0.9988.

Table 2. Results of logistic regression for selected variables in Set 1

Analysis of Penalized Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-12.4444	4.3187	8.3033	0.004
x38	1	1.3209	0.5184	6.4918	0.0108
x39	1	0.9806	0.381	6.6251	0.0101

Source: own calculations

Logit function is of the form $g(x) = -12.4444 + 1.3209\mathrm{x}_{38} + 0.9806\mathrm{x}_{39}$

The probability of default for client with relevant values of x38 and x39 can be evaluated as

$$\pi(x) = \frac{\exp(g(x))}{1+\exp(g(x))}.$$

$\exp(\beta_1) = \exp(1.3187) = 3.75$. This value has an economic interpretation. Namely, the increase of x38 by one unit increases the odds of default almost 4 times.

Model built for Set 2 with 58 variables with Firth amendment provided results presented in Table 3. Only 2 variables were significant. Area under the ROC curve for Set 2 was 0 0.9930, which is very good.

Table 3. Results of logistic regression for selected variables in Set 2

Analysis of Penalized Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.732	0.9399	51.2996	<0.0001
x38	1	0.2424	0.0723	11.2429	0.0008
x39	1	1.0281	0.165	38.8434	<0.0001

Source: own calculations

Model built for Set 3 with 49 variables with Firth amendment provided results presented in Table 4. Only 3 variables were significant. For Set 3 AUC was 0.9859.

Table 4. Results of logistic regression for selected variables in Set 3

Analysis of Penalized Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.1704	0.7583	89.42	<0.0001
x38	1	0.3636	0.0799	20.7128	0.0001
x39	1	0.4793	0.126	14.4618	<0.0001
x40	1	0.4716	0.1065	19.6009	<0.0001

Source: own calculations

Feature selection by machine learning methods

Preliminary selection left quite large numbers of variables for each set. Therefore, we have performed regularization methods to decrease the numbers of features for further selection. We have also calculated feature importance for preliminary selected sets of features, using Random Forests, Gradient Boosting, XGB and Neural Networks.

Set 1

L1 selection (LASSO) with C=2 distinguished variables x3, x27, x34, x38, x39, x53, var11 and L2 (Ridge) with C=0.01 distinguished: x8, x23, x24, x25, x29, x34, x43, x44, x46, x48, x52, x55, x56, x57, x59, var1, var2, var4, var5, var7, var8, var11, var12, var13, var14, var15. The hyperparameter C has been tuned. For a selected value C=1.5 the following features were extracted: x3, x8, x12, x23, x24, x27, x29, x34, x38, x39,x51, x52, x59, var3, var10, var11, var12, var13,var14. The features that appear in each selected set are x38 and x39.

Set 2

L2 selection with C = 0.01 distinguished the following set of features x8, x20, x23, x24, x25, x30, x34, x39, x40, x43, x44, x51, x54, x56, x57, x59, var4, var7, var9, var11, var12, var13. L1 selection with C=0.1 distinguished x8, x23, x34, x38, x39, x40, x54, x57, x59.

Set 3

L2 (Ridge regularization) with hyperparameter C=1.5 distinguished the following features: x3, x8, x13, x23, x28, x38, x39, x40, x54, var11, var13. L1 with C = 0.01 selected the following features: x23, x34, x38, x39, x40, x43, x44, x54, x57, x59. The features that appear in each selected set are x38, x39 and x40.

We have performed recursive feature elimination for Random Forests. The number of features to be slected was set to 5. The results are shown in Table 5. Feature x38 and x39 appear in each selected set.

Table 5. Features selected by Recursive feature elimination

	Set 1	**Set 2**	**Set 3**
RFE	x34, x38, x39, x52, var3	x38, x39, x40, x46, x54	x23, x38, x39, x40, x54

Source: own calculations

Features distinguished by calculating feature importance in the most popular machine learning algorithms are presented in Table 6. Feature x38 and x39 appear in each set. Additionally, x40 appears in each set for Set 2 and Set 3.

Table 6. Features distinguished by calculating feature importance

Method	**Set1**	**Set2**	**Set3**
RF	x54, x52, x53, x42, x39, x38	x42, x38, x39, x40	x42, x39, x38, x40
GB	x53, x39, x38	x38, x39, x40, x44, x46	x38, x39, x40
XGB	x38, x39	x38, x39, x40, x44, x46	x40, x39, x19, x44, x38

Source: own calculations

Table 7. First 10 features distinguished by calculating relative feature importance in NN

	Set 1		Set2		Set3	
	Name	Relative Importance	Name	Relative Importance	Name	Relative Importance
1	x39	1.0	x39	1.0	x40	1.0
2	x38	0.77	x40	0.75	x38	0.95
3	x19	0.35	x38	0.73	x39	0.85
4	x5	0.27	x18	0.31	x67	0.18
5	x27	0.26	x27	0.29	x34	0.17
6	x18	0.25	x7	0.22	x52	0.15
7	x9	0.24	x5	0.16	x48	0.15
8	x20	0.23	var11	0.16	x3	0.12
9	x47	0.19	var9	0.16	var11	0.12
10	var11	0.19	x3	0.15	x60	0.12

Source: own calculations

Each of the machine learning methods used distinguished different variables, although some variables, e.g., x38 and x39, are repeated in each ranking. It is worth stressing that machine learning methods also distinguished simulated variables (var), although they appear at the end of rankings.

Finally, selected machine learning models performance in terms of accuracy was compared. We have applied Logistic Regression (LR), Random Forests (RF), Gradient Boosting (GB), XGBoost, AdaBoost (AB) and Extra Trees (ET). The results are shown in Table 8. The performance of various models is different on each data set. There is no best model for each set, although ExtraTrees (ET) have the best accuracy for Set 1 and Set 3. One can however notice, that Logistic Regression treated as a machine learning model exhibits the worst performance in all cases.

Table 8 Classification results in terms of accuracy of methods performed for various sets of variables

	Set 1		Set 2		Set 3	
Variables	x34, x38, x39		x38, x39, x40, x46, x54		x23, x38, x39, x40, x54	
	Accuracy	STD	Accuracy	STD	Accuracy	STD
LR	0.935	(0.1026)	0.9525	(0.0425)	0.96	(0.0236)
RF	0.985	(0.0229)	0.9775	(0.0284)	0.9813	(0.0151)
GB	0.98	(0.0245)	0.9675	(0.0372)	0.98	(0.0139)
XGB	0.97	(0.04)	0.9725	(0.0261)	0.9788	(0.0148)
AB	0.985	(0.0229)	0.97	(0.0312)	0.9738	(0.0181)
ET	0.99	(0.03)	0.9725	(0.0236)	0.9825	(0.0127)

Source: own calculations

CONCLUSIONS

One of the basic elements of building models is the selection of appropriate independent variables.

Independent variables are selected to represent expected influences based on: theory (often relatively weak), previous research, and local context (in time and space). In the statistical approach, the main emphasis is placed on the sign, magnitude and statistical significance of the weights for the independent variables. Algorithmic models require different approaches.

The article presents the results of the selection of variables used in practice and dedicated to specific types of models. The work was carried out on a relatively large data set to avoid problems related to the so-called low power effects.

Feature importance computed on large feature sets produced stable results. The number of selected features is small and some of them are repeated in different analyses. It can be stated that up to a certain limit value of the so-called of practical importance, various selection algorithms correctly identify relationships between independent variables and the target variable. The situation is different in the case of a weak relationship, where there is a problem of the so-called multiplicity of data models. This means that prediction accuracy becomes more robust as the set of independent variables changes. Unfortunately, this property of the models makes it much more difficult to correctly interpret the results from a substantive point of view.

Based on the results obtained, it can be concluded that there are large differences between the models. It can therefore be concluded that attempts to reconcile results between different analytical approaches must be carried out very carefully and should take into account the fact that the definitions of "variables of significance" are strongly dependent on the model.

REFERENCES

Adler A. I., Painsky A. (2022) Feature Importance in Gradient Boosting Trees with Cross-Validation Feature Selection. Entropy, 24(5), 687. https://doi.org/10.3390/e24050687

Bag S., Gupta K., Deb S. (2022) A Review and Recommendations on Variable Selection Methods in Regression Models for Binary Data. https://arxiv.org/pdf/2201.06063

Ben Jabeur S., Stef N., Carmona P. (2023) Bankruptcy Prediction using the XGBoost Algorithm and Variable Importance Feature Engineering. Computational Economics, 61, 715-741. https://doi.org/10.1007/s10614-021-10227-1

Breiman L. (2001) Statistical Modeling: The Two Cultures. Statistical Science, 16(3), 199-215.

Dorogush A.V., Ershov V., Gulin A. (2017) CatBoost: Gradient Boosting with Categorical Features Support. Workshop on ML Systems at NIPS 2017.

De Sa C. R. (2019) Variance-Based Feature Importance in Neural Networks. [in:] Kralj Novak P., Šmuc T., Džeroski S. (eds) Discovery Science, Lecture Notes in Computer Science. 11828, Springer, Cham. https://doi.org/10.1007/978-3-030-33778-0_24

Engelmann B., Rauchmeier R. (2011) The Basel II: Risk Parameters. Estimation, Validation, Stress Testing - with Applications to Loan Risk Management. Springer Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-16114-8

Elith J., Leathwick J. R., Hastie T. (2008) A Working Guide to Boosted Regression Trees. Journal of Animal Ecology, 77, 802-813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Firth D. (1993) Bias Reduction of Maximum Likelihood Estimates. Biometrika, 80(10), 27-38. https://doi.org/10.2307/2336755

Gajowniczek K. et al. (2022) HOFS: Higher Order Mutual Information Approximation for Feature Selection in R. SoftwareX, 19, 1-9. https://doi.org/10.1016/j.softx.2022.101148

Hastie T., Tibshirani R., Friedman J. (2008) The Elements of Statistical Learning (2nd ed.). Springer.

Hastie T., Tibshirani R., Wainwright M. (2015) Statistical Learning with Sparsity. The Lasso and Generalizations. New York Chapman & Hall/CRC. https://doi.org/10.1201/b18401

Hopf K., Sascha R. (2021) Filter Methods for Feature Selection in Supervised Machine Learning Applications-Review and Benchmark. arXiv preprint arXiv:2111.12140, 2021

Jia W., Sun M., Lian J. et al. (2022) Feature Dimensionality Reduction: a Review. Complex Intell. Syst., 8, 2663-2693. https://doi.org/10.1007/s40747-021-00637-x

Kohavi R, John G. H. (1997) Wrappers for Feature Subset Selection. Artificial Intelligence, 97(1-2), 273-324. https://doi.org/10.1016/S0004-3702(97)00043-X

Lal T. N., Chapelle O., Weston J., Elisseeff A. (2006) Embedded Methods. [in:] Guyon I., Nikravesh M., Gunn S., Zadeh L. A. (eds) Feature Extraction. Studies in Fuzziness and Soft Computing, 207, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-35488-8_6

Li J., Cheng K., Wang S., Morstatter F. (2017) Feature Selection: A Data Perspective ACM Computing Surveys, 50(6), Article 94, 1-45. https://doi.org/10.1145/3136625

Olden J., Joy M., Death R. (2004) An Accurate Comparison of Methods for Quantifying Variable Importance in Artificial Neural Networks using Simulated Data. Ecological Modelling, 178(3-4), 389-397. https://doi.org/10.1016/j.ecolmodel.2004.03.013

Ostroumova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. (2017) CatBoost: Unbiased Boosting with Categorical Features. NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems December 2018, Pages 6639–6649. https://arxiv.org/pdf/1706.09516

Priyatno A.,Widiyaningtyas T. (2024) A Systematic Literature Review: Recursive Feature Elimination Algorithms. JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer). 9. 196-207. https://doi.org/10.33480/jitk.v9i2.5015

Pudjihartono N., Fadason T., Kempa-Liehr A. W., O'Sullivan J. M. (2022) A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics, 2(927312). https://doi.org/10.3389/fbinf.2022.927312

Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A. (2017) Firth's Logistic Regression with Rare Events: Accurate Effect Estimates and Predictions? Stat Med. 36(14), 2302–2317. https://doi.org/10.1002/sim.7273

Sánchez-Maroño N., Alonso-Betanzos, A., Tombilla-Sanromán, M. (2007) Filter Methods for Feature Selection – A Comparative Study. [in:] Yin, H., Tino, P., Corchado, E., Byrne, W., Yao, X. (eds) Intelligent Data Engineering and Automated Learning - IDEAL 2007. IDEAL 2007. Lecture Notes in Computer Science, 4881, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-77226-2_19

Sauerbrei W., Perperoglou A., Schmid M. et al. (2020) State of the Art in Selection of Variables and Functional Forms in Multivariable Analysis-Outstanding Issues. Diagn Progn Res 4, 3. https://doi.org/10.1186/s41512-020-00074-3

Taylor J., Tibshirani R. J. (2015) Statistical Learning and Selective Inference. Proc Natl Acad Sci U S A, 112(25), 7629-34. https://doi.org/10.1073/pnas.1507583112

Wojtas M., Chen K. (2020) Feature Importance Ranking for Deep Learning. 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. https://doi.org/10.48550/arXiv.2010.08973

Vergara J. R., Estévez P.A. (2014) A Review of Feature Selection Methods based on Mutual Information, Neural Comput & Applic, 24, 175-186. https://doi.org/10.1007/s00521-013-1368-0

Zebari R., Abdulazeez A., Zeebaree D., Zebari D., Saeed, J. (2020) A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction, Journal of Applied Science and Technology Trends, 1(2), 56-70. https://doi.org/10.38094/jastt1224

APPENDIX

Description of variables distinguished in various models

Variable	Description
x3	Ratio of the sum of debit balances at the moment of analysis to the average debit balance in the last 12 months
x5	Linear trend of average monthly payments for the period - 6 months
x7	Increasing trend for average monthly payments for the period - 6 months
x8	Ratio of the sum of balances at the moment of analysis to the average total balance over the last 12 months
x9	Average debit balance over the last 3 months
x13	Average increase in the amount of capital arrears in the last period
x19	Average credit balance over the last 3 months
x20	Average credit balance over the last 6 months
x23	Ratio avg. the amount of overdue capital installments up to avg. amount of the total credit balance (on all customer accounts) in the last month) for - 6 months
x24	Ratio avg. the amount of overdue capital installments up to avg. amount of the total credit balance (on all customer accounts) in the last month) for - 9 months
x25	Ratio avg. the amount of overdue capital installments up to avg. amount of the total credit balance (on all customer accounts) in the last month) for - 12 months
x27	Average time of delay in repayment of installments, determined as the ratio of the sum of days of delay for all installments paid to the number of all installments repaid.) for - 12 months
x28	Average time of delay in repayment of installments, determined as the ratio of the sum of days of delay for all installments paid to the number of all installments repaid.)
x29	Number of overdue accounts as at the date of analysis) for - 6 months
x30	Ratio of the sum of debit balances at the moment of analysis to the average debit balance in the last 12 months
x34	Number of overdue accounts as at the date of analysis) for - 9 months
x38	Sum of amounts from all months of repayments made by the client on credit accounts held by him) for - 3 months
x39	Sum of amounts from all monthly repayments made by the client on account of credit accounts held by him) for - 6 months
x40	Sum of all monthly repayments made by the client on credit accounts held by him/her) for - 9 months
x42	Sum of all monthly repayments made by the client on credit accounts held by him)
x43	Sum of amounts from all months of repayments made by the client on credit accounts held by him) for - 1 month
x44	Sum of amounts from all months of repayments made by the client on credit accounts held by him) for - 1 month
x46	Sum of amounts from all months of repayments made by the client on credit accounts held by him) for - 1 month
x47	The sum of interest arrears at the moment of analysis
x48	Loan repayment ratio, determined as the ratio of the value of all repaid loans to the value of all loans taken/granted) for - 12 months
x51	Sum of the amounts of all overdue repayments incurred on all customer credit accounts in the last month) for - 3 months

Variable	Description
x52	Sum of the amounts of all overdue repayments incurred on all customer credit accounts in the last month) for - 6 months
x53	Total amounts of all overdue repayments incurred on all customer credit accounts in the last month) for - 9 months
x54	Sum of the amounts of all overdue repayments incurred on all customer credit accounts in the last month) for - 12 months
x55	Total amounts of all outstanding repayments incurred on all customer credit accounts during the last month)
x56	Sum of the amounts of all overdue repayments incurred on all customer credit accounts in the last month) for - 1 month
x57	Sum of the amounts of all overdue repayments incurred on all customer credit accounts in the last month) for - 1 month
x59	Sum of the amounts of all overdue repayments incurred on all customer credit accounts in the last month) for - 1 month
x60	Increasing trend in average monthly unused limits on all short-term loans over the last 12 months
x66	Amount of all outstanding receivables falling within the range (over 90 days) (summing over all customer credit accounts))
x67	Average monthly unused limit on all short-term loans over the last 6 months

Artificially created variables have been added to business variables: var1 - var15: independent variables with uniform distribution