ORIGINAL PAPER

DOI: 10.26794/2587-5671-2023-27-1-103-115
UDC 336.7(045)
JEL G21, G29

Combined Feature Selection Scheme for Banking Modeling

S.V. Afanasyev[a], D.M. Kotereva[b], A.A. Mironenkov[c], A.A. Smirnova[d]
[a, b, d] National Research University Higher School of Economics, Moscow, Russia;
[a, b, d] Renaissance Credit Bank, Moscow, Russia;
[c] Lomonosov Moscow State University, Moscow, Russia

ABSTRACT

Machine learning methods have been successful in various aspects of bank lending. Banks have accumulated huge amounts of data about borrowers over the years of application. On the one hand, this made it possible to predict borrower behavior more accurately, on the other, it gave rise to the problem a problem of data redundancy, which greatly complicates the model development. Methods of feature selection, which allows to improve the quality of models, are apply to solve this problem. Feature selection methods can be divided into three main types: filters, wrappers, and embedded methods. Filters are simple and time-efficient methods that may help discover one-dimensional relations. Wrappers and embedded methods are more effective in feature selection, because they account for multi-dimensional relationships, but these methods are resource-consuming and may fail to process large samples with many features. In this article, the authors propose a combined feature selection scheme (CFSS), in which the first stages of selection use coarse filters, and on the final — wrappers for high-quality selection. This architecture lets us increase the quality of selection and reduce the time necessary to process large multi-dimensional samples, which are used in the development of industrial models. Experiments conducted by authors for four types of bank modelling tasks (survey scoring, behavioral scoring, customer response to cross-selling, and delayed debt collection) have shown that the proposed method better than classical methods containing only filters or only wrappers.

Keywords: feature selection; machine learning; feature importance; filters; wrappers; embedded methods

For citation: **Afanasyev S.V., Kotereva D.M., Mironenkov A.A., Smirnova A.A. Combined feature selection scheme for banking modeling. *Finance: Theory and Practice*. 2023;27(1):103-115. DOI: 10.26794/2587-5671-2023-27-1-103-115**

INTODUCTION

Machine learning methods have been successful in various aspects of bank lending. Huge amounts of data allow more accurate to predict the behavior of the borrower, while is causing the problem of data redundancy, which complicates the development of models and can lead to unsatisfactory results. To solve this problem, various methods of feature selection were proposed [1]. The basic concept of these methods is to reduce the dimension of the feature space by excluding redundant features.

The methods of feature selection proposed in the scientific literature are divided into three types: filter methods, wrapper methods and embedded methods [2].

Most of the proposed methods in the scientific literature are tested on open repositories that contain either few observations (a few dozen to several thousand), or few features (a few dozen) [3–5]. In practice, bank modeling uses samples, that are orders of magnitude more scientific databases, and include from several hundred thousand to several million observations and from several hundred to several thousand features. On such samples, the methods proposed in the studies either do not produce a declared result or work for a very long time. To solve these problems, we offer the method of Combined Feature Selection Scheme (CFSS), which is a hybrid multi-stage selection scheme, where filters are used in the first stages and wrappers in the subsequent stages. As filters we use methods for cleaning data, feature stability check, feature correlation with target variable, cross-covariance matrix [6] and VIF analysis [7], as wrappers — permutation method

based on a Random Forest [8] and evaluation of p-value features using the Backward Elimination method [9].

Our proposed method was tested on four samples for different banking tasks: prediction of the probability of a credit overdue at the time of application (Application PD), prediction of the probability of future credit overdue during life of the loan (Behavioral PD), evaluation of client response to advertising (CRM PTB) and estimate the transition probability of a credit overdue in a later month (Collection Allocation).

The results showed that the CFSS method works well on large high-dimensional samples. We have also demonstrated that the CFSS method achieves a higher generalization ability of the models through a flexible combination of filters and wrappers than the non-filter wrapper method. Additional experiments have shown that the CFSS method is ten times faster than classical feature sampling methods, which is an important advantage of the method in industrial applications.

REVIEW OF THE LITERATURE

Filters

The simplest feature selection methods include filters that allow the selection of features independently of the model being developed.

The selection of variables by the matrix of correlation (CFS) allows the assessment of subsets of features, based on the hypothesis that good subsets contain features that are not correlated with each other but strongly correlated with the target variable. The simplest way to highlight strongly correlated features is to build matrix of pair correlations features. This approach has been widely adopted in practice. The advantages of the method include simplicity of implementation and interpretation. Disadvantages of the method include sensitivity to data quality (emissions, errors, etc.), as well as inability to identify multi-factor relationships.

Principal Component method (PCA), proposed by Karl Pearson in 1901 [10] and still a popular method in applied problems, reduces dimension by computing the main component of the feature matrix and then reducing the dimension of the matrix through its singular decomposition [11]. Among the advantages of the method can be noted the simplicity of its implementation. The disadvantages of PCA include scale sensitivity, difficulty in selecting cut-off score for the main components, and the fact that PCA does not take into account the target variable, so that the main components may not be informative.

Despite these disadvantages, filters are actively used in practice and are still the subject of scientific research. Zhang and co-authors [12] use Welch's t-test to develop algorithms for early computer diagnosis of Alzheimer's disease

Roffo and Melzi [5] propose a method for selecting features based on the analysis of the graph, where the vertices of the graph are the investigated features, and the edges — are the strength of the connection between the features. The authors assume that the eigenvector will contain rank-by-importance features with the maximum main component in the adjacency matrix of the graph. If the coefficients of linear correlation between features are used as a link function, then the adjacency matrix of the graph becomes a standard correlation matrix.

Wrappers

Among wrapper methods, the most popular was the stepwise regression methods: Forward Selection method [13], Backward Elimination method [13] and Stepwise method [14].[1] Despite its simplicity and effectiveness, stepwise regression methods have been criticized in the scientific community [15].

In the scientific literature also, much attention is paid to metaheuristic optimi-

[1] SAS Institute Inc. (1989) SAS/STAT User's Guide, Version 6, Fourth Edition, Vol. 2, Cary, NC. URL: https://www.scirp.org/reference/ReferencesPapers.aspx? ReferenceID=1542754 (accessed on 07.02.2023).

zation algorithms for feature selection, which include: Particle Swarm Optimization (PSO) [16], Grey Wolf Optimization (GWO) [14], Genetic Algorithm (GA) [17, 18], Bee Swarm Optimization (BSO) [19] and etc. Shen and Zhang proposed an Improved Two-Step Gray Wolf Optimization (IGWO) [3], where at the first stage the authors propose to use the nested regularization method LASSO, on the second — Grey Wolf Optimization method (GWO). Basak and co-authors have proposed the Reinforced Swarm Optimization (RSO) [4] wrappers method, which is an improved Bee Swarm Optimization algorithm, where instead of BSO optimization the reinforcement learning approach is used. Feature selection methods based on metaheuristic algorithms are widely used to select a good approximation in various complex optimization problems, but they do not always provide the best solution because that the training of the final models can be done with other machine learning algorithms for which the selected features may not be optimal.

Among other effective wrappers there are permutation methods based on random forest [18, 20]. In permutation methods, evaluated features are not removed from the sample, i.e. the feature space remains unshifted. Important advantages of permutation methods based on random forest algorithms include the possibility of obtaining unbiased importance estimates using randomized trees (unlike the gradient boosting, where the trees are dependent and the evaluation is biased). The disadvantages of these methods include high computational complexity, which limits the applicability of these methods to large, high-dimensional samples. Celik and co-authors [8] tried to solve this problem by proposing the permutation method New Approach, which showed high efficiency in working speed on large samples. However, the authors have demonstrated that permutation methods do not work well on high-dimensional samples when the number of features is counted several thousand or more.

Embedded Methods

Regularization refers to the embedded methods group where feature selection becomes part of the model building process. In the logistic regression, which became the banking standard [21], the most common regularization methods are L1 (LASSO) [22] and L2 (Ridge) [23] (Tikhonov regularization [24]). The general concept of regularization is to add a penalty element to a functional error that punishes the model for excessive complexity. Regularization of L1 allows to nullify part of weight regression coefficients, and regularization of L2 limits their norm [25]. L1 regularization has a number of disadvantages and does not work well on high-dimensional data with few observations. Zou and Trevor [26] suggest circumventing these limitations with the Elastic Net approach — a combination of L1 and L2 methods.

COMBINED FEATURE SELECTION SCHEME

Pros and cons described in the scientific literature impose limitations on the applicability of the proposed methods of feature selection to practical business tasks. If you need fast methods with low computing requirements, filters will be the most optimal. For higher quality models, wrappers and embedded methods should be used, which may require more computing power. It is important to note that research results may not be reproduced in practice on large, high-dimensional data. These problems motivated us to develop a method of combined feature selection that includes 10 steps of data processing (*Fig. 1*).

Data Quality Analysis

Data quality — a generalized concept reflecting the degree of suitability of data to solve of certain task. Among methods of data quality analysis can be distinguished:

1) Exploratory Data Analysis (EDA) [27] — to identification basic properties of data, to

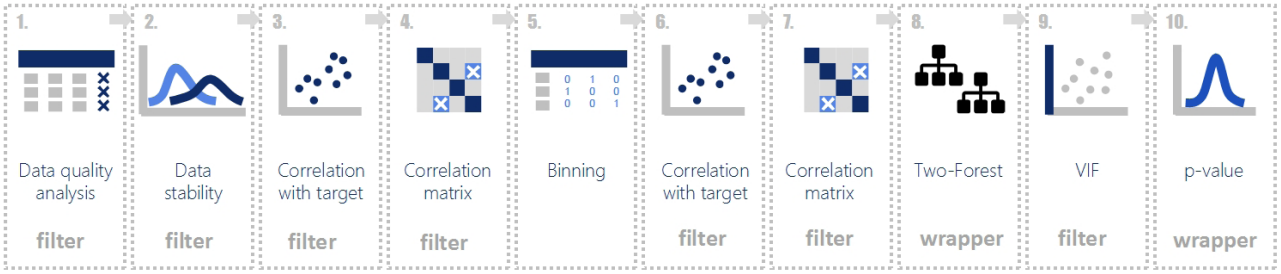

Fig. 1. **Combined feature selection scheme**

Source: Compiled by the authors.

find general patterns, analysis of distributions, emissions, etc.;

2) Analysis of omissions and incomplete data — is conducted using statistical indicators such as the number of non-empty observations, omissions, minimum and maximum values, median, modal value, standard deviation, quantili, etc.;

3) Analysis of anomalies — statistical and expert analysis of the reasons for the occurrence of observations beyond the acceptable range of the variable. The main working methods with anomalies are reduced to the construction of the distribution according to the observed variable and the subsequent definition of thresholds values in the "end" distribution. Alternative methods of working with anomalies are also used, such as monotonic transformation of variables (logarithmic, etc.), calculation of z-score, etc.[2]

Analysis of Data Stability and Continuity

Statistical algorithms depend on continuous and stable data. The reason for the instability in the data can be changes in the bank's business processes, legislation, customer behavior, data formats, etc.

Before evaluating stability, all string variables must be converted to numeric format using LabelEncoder[3] (omissions are replaced with unique numeric value). To estimate the stability of features it is necessary to calculate the divergence between distributions built on different time periods. To do this, stability is assessed for different periods:

1) *large periods:* the sample is divided into equal sub-samples with a large interval (for example, by half-years), after which on these sub-samples in pairs the distributions of features on the principle "each with each" (*Fig. 2a*);

2) *small periods:* the total sample is divided into equal small sub-samples (for example, monthly), after which the features by contiguous periods distributions are compared in pairs (*Fig. 2b*).

In the CFSS scheme, three values are calculated and averaged to calculate the divergence between sub-samples: S- statistics [28], population stability index (PSI) [29] and Kolmogorov-Smirnov statistics (KS) [30].

This method allows to detect both long-term changes features (instability over large periods) and frequent short-term changes (instability for small periods).

Correlation of Features with the Target Variable

Correlation analysis of features with target variable allows to select features that strongly influence target variable. At the same time, this method does not take into account complex dependencies between the features, so it can be attributed to "rude" filter methods that can be used for the primary selection of features.

Correlation analysis methods depend on the type of target variable and the type researched feature.[4]

[2] Understanding Statistics. Graham J.G. Upton, Ian T. Cook. Oxford University Press, 1996. URL: https://books.google.ru/books?id=vXzWG09_SzAC&printsec=frontcover&hl=ru#v=onepage&q&f=false (accessed on 30.01.2023).

[3] URL: scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html (accessed on 30.01.2023).

[4] Aivazyan S.A., Mkhitaryan B.S. Applied statistics. Basics of econometrics. Textbook for universities. In 2 vol., 2nd edition. Vol. 1. Probability theory and applied statistics. Moscow: IUNITI-DANA; 2001. 656 p.

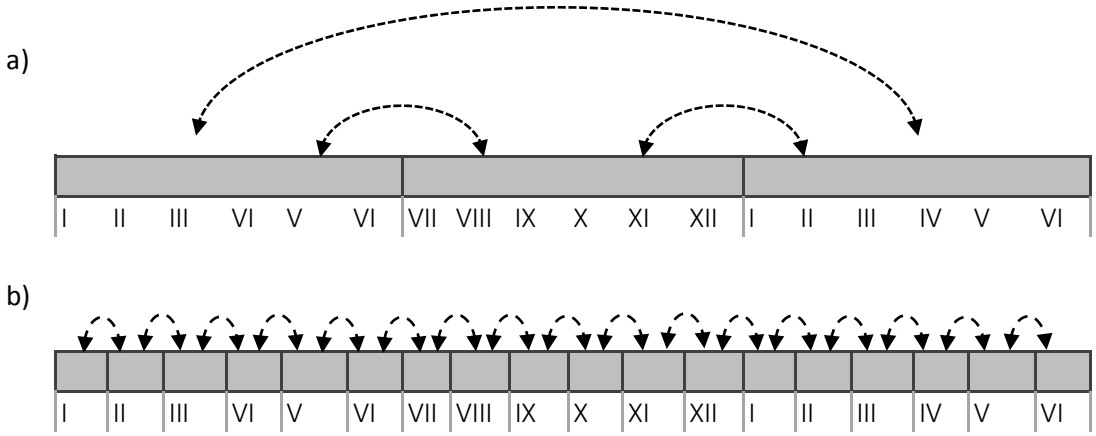

Fig. 2. **An Example of a Split into Periods to Estimate the Stability of Feature a) large Periods; b) Small Periods**

Source: Compiled by the authors.

Note: Months are indicated by Roman numerals.

Features, that do not pass test on correlation value with target variable, are excluded from further analysis. Significance thresholds are set as heuristic or experimentally selected.

In the combined CFSS scheme, features are checked for correlation with the target variable twice — before binarization of features and after (step 6, *Fig. 1*).

Matrix of Correlation

Highly correlated features can be detected using a correlation matrix (CFS), which has the form:

$$R_x = \begin{pmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_n} \\ r_{x_2x_1} & 1 & \dots & r_{x_2x_n} \\ \dots & \dots & \dots & \dots \\ r_{x_nx_1} & r_{x_nx_2} & \dots & 1 \end{pmatrix}, \quad (1)$$

where $r_{x_ix_j}$ — correlation between i and j features.

Pearson correlation is calculated for continuous features and Spearman's correlation for categorical and binary features.

In the combined CFSS scheme, feature selection with the correlation matrix is performed twice — before and after feature binarization (step 7, *Fig. 1*). Threshold values for feature selection are set as heuristics and depend on the type of task. For our experiments we set the following thresholds: 90% for the stage "before binarization of features" (weak filtering) and 70% for the stage "after binarization of features" (strong filtering).

Dummy-Coding of Categorical Variables

After the feature selection by primary filters, it is necessary to convert the categorical features into binary variables for their possible use in regression algorithms. For logistic regression, the dummy-coding procedure [31] is used by full rank method when one of the categories is removed. Thus, after dummy-coding the categorical variable is $k - 1$ binary variable, where k — is the number of categories in the original feature.

After transformation of categorical feature for each new binary variable the number of observations of "positive" class will be calculated. All binary variables for which the number of observations of the "positive" class is less than the specified threshold of significance is combined into one category. For other types of target variables, the number of observations in the category is considered. Significance threshold is set as heuristic or experimentally selected. For our experiments, we set the threshold at 10.

Two-Forest Method

After primary feature filtering, CFSS uses wrapper methods that consider complex relationships between features. Random forest selection methods are the most accurate wrappers [18, 20]. To select a random forest method, need to evaluate the importance of each features using one of two approaches:

1. Importance by reducing heterogeneity:

1) for each tree in a random forest, the sum of the decreases in the heterogeneity of all branches associated with this variable is calculated;

2) the total amount of heterogeneity decreases is divided by the total number of trees;

3) steps (1) and (2) are repeated for all variables.

The desired importance of a feature — is the frequency with which a variable is used as a predictor of branching.

2. Importance based on reduced quality forecasting in case of random permutation:

1) Random forest model is trained;

2) an error is calculated on the test/OOB multiple;[5]

3) variable (or group of variables) is fixed and its values are randomly rearranged on test/OOB multiple;

4) calculates the difference between the error on the original multiple and the error on the multiple with the permutation.

The calculated error subtraction is the permutation importance of the variable.

The CFSS scheme uses an adapted New Approach [8] method, which we called Two-forest.[6] The general concept of the Two-forest method is evaluation the importance of features as quality forecasting at random permutation:

$$VI_j = P\left(Y \neq f\left(X_1,\ldots,X_j^*,\ldots,X_p\right)\right) - P\left(Y \neq f\left(X_1,\ldots,X_j,\ldots,X_p\right)\right). \quad (2)$$

1. Adapted Two-forest method works according to the following algorithm:

2. The training sample is representative of two equal parts.

3. Each sub-sample are training a random forest.[7]

4. Quality is assessed at sub-sample on which the model was not trained.

5. Each variable is randomly mixed and the result for each of the two models on the sub-samples is considered.

6. Calculates the subtraction between the baseline value obtained in step 3 and the new value.

7. The importance of the variable is calculated as the average of the importance value on two sub-samples.

The p-value value is calculated for the resulting value:

1) observations with negative values of importance are selected;

2) zero-value observations are selected;

3) negative values of importance are multiplied by (–1);

4) vectors obtained from steps (1)–(3) are concatenated;

5) a cumulative distribution is constructed for the resulting vector;

6) on the received distribution p-value is calculated.

8. Variables with p-value below a given threshold are selected. The following heuristics are possible:

1) divide the importance by the average value of the baseline, those variables whose change exceeds the specified threshold are selected;

2) sort the variables by importance values and select the first N variables (the number N

[5] OOB (Out-of-Bag) — quality assessment for each observation only for those trees of the ensemble that were not trained on this observation (i.e. using objects that were not part of the training sample for each base tree).

[6] Authors called the method "Two-forests" because in this approach learns two random forests at once.

[7] Different algorithms are used depending on the type of target variable: Random Forest Classifier — for the binary target variable; Random Forest Regressor — for continuous target variable.

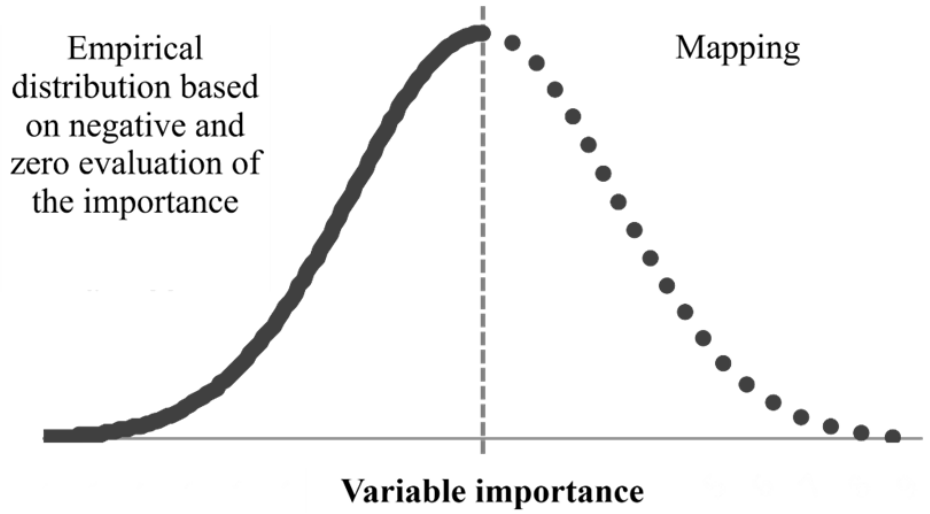

Fig. 3. **Example of Constructing an F Distribution Based on Irrelevant Features (i.e. with Negative or Zero Importance Evaluations)**

Source: Compiled by the authors.

is experimentally selected), the value p-value of the selected variables should be less than 10%.

VIF Analysis

Another approach to reduce multicollinearity between features is based on the estimation of indicator VIF (Variance Inflation Factor) [7]. To calculate this indicator, it is necessary to construct a linear regression for each explanatory factors (as a target variable) from all other features. Selection by VIF analysis is based on the following algorithm:

1. For each feature X_i is training linear regression, in which X_i is a function of all other features:

$$X_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_j, \quad i \neq j, \tag{3}$$

where β_0 — is the member of regression;

k — total number of features (including analysed).

2. VIF coefficient for feature X_i is calculated:

$$VIF_i = \frac{1}{1 - R_i^2}, \tag{4}$$

where R_i^2 — coefficient determinant of regression based on step 1.

3. Evaluation of obtained VIF values is carried out, where a general empirical rule is applied: features with VIF > 10 refer to multicollinear [32]. A feature with a maximum VIF value is removed from the list of multicollinear features.

4. Steps 1–3 are iteratively repeated until the maximum value of VIF for the remaining features is less than or equal to 10.

Statistical Significance of Features

The final step in the CFSS scheme is to check the statistical significance of the features based on the likelihood ratio test.

The procedure estimate of statistical significance features by a test of the likelihood ratio boils down to the verification of the null hypothesis of the value of feature through the evaluation of the probability ratio statistics. For a model with a parameter vector β it is necessary to test the $H_0 : g(\beta) = 0$, hypothesis with sample data, where $g(\beta)$ — collection (vector) of some parameter functions. To test the null hypothesis, the likelihood functions of the full model (i.e. trained on all n features) are compared to the shortened model without the feature being tested (trained on $n - 1$ remaining features). To do this, we calculate the likelihood ratio test:

$$LR = 2 \cdot (L_l - L_s) = 2 \cdot \ln \frac{L_l}{L_s}, \tag{5}$$

where L_l — the value of logarithmic function likelihood of full model;

Table 1

Characteristics of Samples for Banking Modeling Tasks

DataSet	Period of train and test samples	Out-of-time period (OOT)	Observations, amount	Features, amount	Minority class Percentage (bad-rate), %
PTB (CRM)	01.11.2019–30.01.2020	01.02.2020–28.02.2020	545 963	1222	1.28
Behavioral PD (Scoring)			1 195 466	1087	13.88
Application PD (Scoring)			793 080	423	3.60
Allocation (Collection)	01.06.2018–30.04.2019	01.05.2019–30.06.2019	256 220	162	37.19

Source: Compiled by the authors.

L_s — the value of logarithmic function likelihood of shortened model.

LR statistics under the null hypothesis have chi-square distributions with q degrees of freedom — $\chi^2(q)$, where q — number of restrictions (number of excluded features). If the value of this statistic is greater than the critical value of the distribution at a given level of significance, then the excluded feature is considered relevant and the full model is preferred. Otherwise, the deleted variable is recognized insignificant.

The p-value threshold is set as a heuristic or experimentally selected. In the tested CFSS scheme, the value level p-value was set to 0.05.

EXPERIMENTAL EQUIPMENT

Data

Data from a large Russian bank were used for experiments. Comparison selection methods were performed on four datasets for banking binary classification tasks:

1. *CRM: PTB (probability to bay)* — model of estimation of client's response to cross-sell loan offer.
2. *Scoring: Application PD (CASH)* — a model for estimating the probability of default at the lending stage (cash loans for individuals).
3. *Scoring: Behavioral PD (CASH)* — a model for estimating the probability of default over the lifetime of a loan, using behavioral information about the client's previous credit payments. This model allows to assess the level of credit risk on the loan portfolio for reserve and capital formation in accordance with the requirements of international financial reporting (IFRS 9) and on the basis of domestic ratings (IRB, Basel II).
4. *Collection: Allocation* — a model for estimating the probability of overdue on a loan in the later month of the payment schedule.

Characteristics of samples are presented in *Table 1*.

Experiment

The second experiment was conducted to compare the classic selection methods popular in banking practice and the CFSS scheme using the Two-forest method. Comparison between three selection schemes:

1. *Gini Scheme* — in this scheme at step 8 (*Fig. 1*) instead of Two-forest method the selection of features using Gini estimates was

Table 2

Comparison of Three Feature Selection Schemes: Gini, Forward, CFSS

Scheme	Models	Features, amount	Gini			
			LogReg		LightGBM	
			Test	OOT	Test	OOT
Gini Scheme (GS)	PTB (CRM)	1222	0.4157	0.4312	0.4380	0.4480
	Behavioral PD (Scoring)	1087	0.6843	0.6400	0.6904	0.6493
	Application PD (Scoring)	423	0.4051	0.3980	0.4251	0.4179
	Allocation (Collection)	162	0.6048	0.6075	0.6499	0.6494
Forward Scheme (FS)	PTB (CRM)	1222	0.4259	**0.4369**	0.4302	0.4481
	Behavioral PD (Scoring)	1087	**0.6907**	**0.6466**	**0.7068**	**0.6705**
	Application PD (Scoring)	423	**0.4164**	**0.4067**	0.4356	0.4203
	Allocation (Collection)	162	**0.6143**	0.6041	0.6418	0.6436
CFSS	PTB (CRM)	1222	**0.4332**	0.4340	**0.4401**	**0.4527**
	Behavioral PD (Scoring)	1087	0.6881	0.6439	0.7050	0.6682
	Application PD (Scoring)	423	0.4093	0.4051	**0.4390**	**0.4290**
	Allocation (Collection)	162	0.6111	**0.6085**	**0.6500**	**0.6507**
	Δ Gini					
Difference: (CFSS – GS)	PTB (CRM)	1222	**0.0174**	**0.0028**	**0.0021**	**0.0047**
	Behavioral PD (Scoring)	1087	**0.0038**	**0.0040**	**0.0146**	**0.0189**
	Application PD (Scoring)	423	**0.0042**	**0.0071**	**0.0139**	**0.0111**
	Allocation (Collection)	162	**0.0062**	**0.0011**	**0.0001**	**0.0013**
Difference: (CFSS – FS)	PTB (CRM)	1222	**0.0073**	-0.0030	**0.0100**	**0.0046**
	Behavioral PD (Scoring)	1087	–0.0025	–0.0027	–0.0018	–0.0023
	Application PD (Scoring)	423	–0.0072	–0.0016	**0.0034**	**0.0087**
	Allocation (Collection)	162	–0.0032	**0.0044**	**0.0082**	**0.0071**

Source: Compiled by the authors.

applied (Gini > 5%) for single factor logistic regressions (all other stages of selection of the general scheme remained unchanged).

2. *Forward Scheme* — in this scheme in step 8, feature selection using the method of Forward Selection was applied.

3. *CFSS* — combined selection using the Two-forest method (*Fig. 1*).

The methods listed were estimated as part of a 10-stage combination selection scheme, in order not to compare the obviously weak Gini and Forward methods with a strong Two-forest method.

As part of the second experiment, the time of the Forward and Two-Forest methods was also estimated.

RESULTS

The objective of the experiment was to compare the CFSS scheme with industry standard methods. Combined selection schemes were compared (*Fig. 1*) with the addition of three different selection methods in step 8: Gini (banking standard), Forward (banking standard) and Two-Forest (CFSS — our approach).

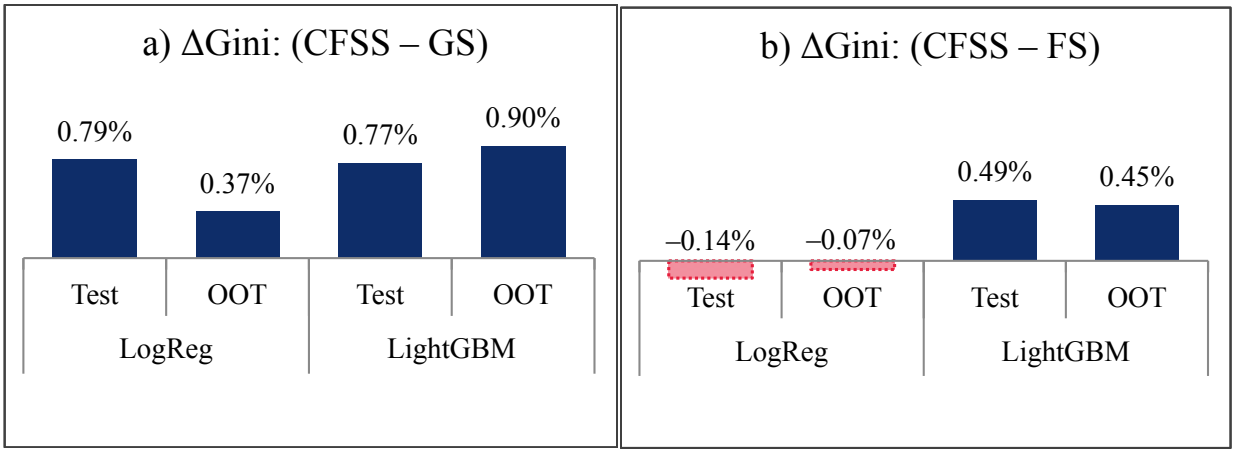

Fig. 4. **Comparison of Gini Scheme (GS), Forward Scheme (FS) and CFSS: a) Gini Difference between CFSS and GS Models; b) Gini Difference between CFSS and FS Models**

Source: Compiled by the authors.

Table 3

Selection Time of the Forward and Two-Forest Methods

Model	Observations, amount (Train)	Features, amount	Time (hh: mm: ss)		x-Times: Forward / 2Forest
			Forward	Two-Forest	
PTB (CRM)	303 220	1222	14:01:28	0:30:48	74x
Behavioral PD (Scoring)	588 385	1087	16:11:04	0:16:36	58x
Application PD (Scoring)	497 063	423	5:34:12	0:15:00	22x
Allocation (Collection)	172 250	162	3:06:40	0:05:50	32x

Source: Compiled by the authors.

The results of experiments (*Table 2*) showed that the Gini scheme lost on quality of Forward and CFSS schemes. On the other hand, the Forward scheme showed better results for logistic regression, and the CFSS scheme — for gradient boosting in LightGBM implementation. This result confirms the thesis that to choose the wrappers you need to consider the type of algorithm for the final model.

A comparison of the average quality difference of the studied models showed that the CFSS scheme lost slightly to the FS scheme (*Fig. 4*) in logistic regression, demonstrating the good stability of the CFSS scheme to the type of algorithm for training the final model. This may be due to the fact that scheme CFSS uses a linear Backward (p-value) method after the nonlinear Two-Forest method, which balances selection towards linear features.

A time comparison of Forward and Two-Forest methods showed that Two-Forest works ten times faster than the Forward method (*Table 3*). In this experiment, the during operation methods were compared only in step 8 of the overall scheme, as all other selection steps were the same.

CONCLUSION

In this article we proposed a scheme of combined CFSS feature selection, in which the first stages are performed feature cleaning and stability checking, the next steps are used correlation filters, allowing to eliminate highly correlated features among themselves, and in the final stages the wrapper methods are applied, which are fine-tuning the scheme and final selection. This selection scheme "from simple to complex" allows to balance the selection and to achieve good results in quality

and speed on large high-dimensional samples.

The results of our experiments showed that the CFSS scheme works well for different types of models (linear and non-linear) and different banking tasks (credit scoring, advertising campaigns, collect overdue debts, etc.) and exceeds the quality of the scheme, containing only filters or wrappers.

Inclusion of multiple wrapper methods in the combined selection scheme allows to control the correctness of each method on the previous selection steps.

Compared to regression approaches, the Two-Forest selection method shows better quality for non-linear models and comparable quality for linear. At the same time, the Two-Forest method scores ten times faster than the regression methods.

The combined feature selection scheme can be fully automated by integrating it into the overall pipeline development models in the bank. This allows the development of models in the mode "End-to-End", which speeds up the development process and reduces model risks.

It should be noted that the CFSS scheme used a set of fixed threshold metrics defined by an expert. Thus, the CFSS scheme is still metaheuristic when data specificities are not considered at some stages of selection. Heuristics data as well as CFSS methods can be further configured as hyperparameters of the model, which will take into account the specifics of the task and improve the quality of the final models. However, configuring hyperparameters will increase the time complexity of the CFSS scheme. Our future researches will be dedicated to these issues.

ACKNOWLEDGEMENTS

The research was carried out at the expense of a grant from the Russian Science Foundation (project No. 20-68-47030). Lomonosov Moscow State University, Moscow, Russia.

REFERENCES

1. Guyon I., Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning Research*. 2003;3(7–8):1157–1182. DOI: 10.1162/153244303322753616
2. Hamon J. Optimisation combinatoire pour la sélection de variables en régression en grande dimension: Application en génétique animale. Docteur en Informatique Thèse. Lille: Université des Sciences et Technologie de Lille; 2013. 160 p. URL: https://core.ac.uk/download/pdf/51213307.pdf
3. Shen C., Zhang K. Two-stage improved Grey Wolf optimization algorithm for feature selection on high-dimensional classification. *Complex & Intelligent Systems*. 2022;8(4):2769–2789. DOI: 10.1007/s40747–021–00452–4
4. Basak H., Das M., Modak S. RSO: A novel reinforced swarm optimization algorithm for feature selection. arXiv:2107.14199. URL: https://arxiv.org/pdf/2107.14199.pdf
5. Roffo G., Melzi S. Features selection via eigenvector centrality. In: Proc. 5th Int. workshop on new frontiers in mining complex patterns (NFMCP2016). (Riva del Garda, 19 September, 2016). Cham: Springer-Verlag; 2017. (Lecture Notes in Computer Science. Vol. 10312). URL: https://www.researchgate.net/publication/305918391_Feature_Selection_via_Eigenvector_Centrality
6. Hall M.A. Correlation-based feature selection for machine learning. PhD thesis. Hamilton: The University of Waikato; 1999. 198 p. URL: https://www.lri.fr/~pierres/donn%E 9es/save/these/articles/lpr-queue/hall99correlationbased.pdf
7. James G., Witten D., Hastie T., Tibshirani R. An introduction to statistical learning: With applications in R. 8th ed. New York, NY: Springer Science+Business Media; 2017. 440 p. (Springer Texts in Statistics).
8. Janitza S., Celik E., Boulesteix A.-L. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*. 2018;12(4):885–915. DOI: 10.1007/s11634–016–0276–4
9. Magnus Ya.R., Katyshev P.K., Peresetskii A.A. Econometrics. Moscow: Delo; 2004. 576 p. (In Russ.).

10. Pearson K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. 1901;2(11):559–572. DOI: 10.1080/14786440109462720
11. Aivazyan S.A., Bukhshtaber V.M., Enyukov I.S., Meshalkin L.D. Applied statistics. Classification and dimensionality reduction. Moscow: Finansy i statistika; 1989. 607 p. (In Russ.).
12. Zhang Y., Dong Z., Phillips P., Wang S., Ji G., Yang J., Yuan T.-F. Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning. *Frontiers in Computational Neuroscience*. 2015;9:66. DOI: 10.3389/fncom.2015.00066
13. Hocking R.R. The analysis and selection of variables in linear regression. *Biometrics*. 1976;32(1):1–49. DOI: 10.2307/2529336
14. Mirjalili S., Mirjalili S.M., Lewis A. Grey wolf optimizer. *Advances in Engineering Software*. 2014;69:46–61. DOI: 10.1016/j.advengsoft.2013.12.007
15. Flom P.L., Cassell D.L. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. In: Northeast SAS Users Group 2007 (NESUG 2007). (Baltimore, 11–14 November, 2007). URL: https://www.lexjansen.com/pnwsug/2008/DavidCassell-StoppingStepwise.pdf
16. Eberhart R., Kennedy J. A new optimizer using particle swarm theory. In: Proc. 6th Int. symp. on micro machine and human science (MHS'95). (Nagoya, 04–06 October, 1995). Piscataway, NJ: IEEE; 1995:39–43. DOI: 10.1109/MHS.1995.494215
17. Schott J.R. Fault tolerant design using single and multicriteria genetic algorithm optimization. PhD thesis. Cambridge, MA: Massachusetts Institute of Technology; 1995. 201 p. URL: https://dspace.mit.edu/handle/1721.1/11582
18. Karaboga D. An idea based on honey bee swarm for numerical optimization. Technical Report. 2005;(06). URL: https://abc.erciyes.edu.tr/pub/tr06_2005.pdf
19. Altmann A., Toloşi L., Sander O., Lengauer T. Permutation importance: A corrected feature importance measure. *Bioinformatics*. 2010;26(10):1340–1347. DOI: 10.1093/bioinformatics/btq134
20. Hapfelmeier A., Ulm K. A new variable selection approach using random forests. *Computational Statistics & Data Analysis*. 2013;60:50–69. DOI: 10.1016/j.csda.2012.09.020
21. Louzada F., Ara A., Fernandes G.B. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*. 2016;21(2):117–134. DOI: 10.1016/j.sorms.2016.10.001
22. Santosa F., Symes W.W. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*. 1986;7(4):1307–1330. DOI: 10.1137/0907087
23. Hilt D.E., Seegrist D.W. Ridge: A computer program for calculating ridge regression estimates. USDA Forest Service Research Note. 1977;(236). URL: https://ia803007.us.archive.org/23/items/ridgecomputerpro236hilt/ridgecomputerpro236hilt.pdf
24. Tikhonov A.N. Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics. Doklady*. 1963;(4):1035–1038. (In Russ.: *Doklady Akademii nauk SSSR*. 1963;151(3):501–504.).
25. Vorontsov K.V. Lectures on regression recovery algorithms. December 21, 2007. URL: http://www.ccas.ru/voron/download/Regression.pdf (In Russ.).
26. Zou H., Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*. 2005;67(2):301–320. DOI: 10.1111/j.1467–9868.2005.00503.x
27. Bruce P., Bruce A. Exploratory data analysis. In: Practical statistics for data scientists: 50 essential concepts. Beijing: O'Reilly Media; 2017;1–46. (Russ. ed.: Bruce P., Bruce A. Razvedochnyi analiz dannykh. Prakticheskaya statistika dlya spetsialistov Data Science. St. Petersburg: BHV-Peterburg; 2018:19–58.).
28. Afanasiev S., Smirnova A. Predictive fraud analytics: B-tests. *Journal of Operational Risk*. 2018;13(4):17–46. DOI: 10.21314/JOP.2018.213
29. Lin J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*. 1991;37(1):145–151. DOI: 10.1109/18.61115

30. Kolmogorov A. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*. 1933;4:83–91.
31. Harris D., Harris S. Digital design and computer architecture. 2nd ed. San Francisco, CA: Morgan Kaufmann; 2012. 720 p.
32. Kutner M.H., Nachtsheim C.J.; Neter J. Applied linear regression models. 4th ed. New York, NY: McGraw-Hill/Irwin; 2004. 701 p.

ABOUT THE AUTHORS

Sergey V. Afanasiev — Master's student, National Research University Higher School of Economics, Moscow, Russia; Vice President, Head of the Statistical Analysis Department, Bank "Renaissance Credit", Moscow, Russia
https://orcid.org/0000-0001-5119-507X
svafanasev@gmail.com

Diana M. Kotereva — Master's student, National Research University Higher School of Economics, Moscow, Russia; Head of Modeling and Operational Analysis Department Bank "Renaissance Credit", Moscow, Russia
https://orcid.org/0000-0001-6102-0222
dmkotereva@edu.hse.ru

Alexey A. Mironenkov — Senior Lecturer at the Department of Econometrics and Mathematical Methods of Economics Moscow School of Economics, Lomonosov Moscow State University, Moscow, Russia
https://orcid.org/0000-0001-5754-8825
Corresponding author:
mironenkov@mse-msu.ru

Anastasiya A. Smirnova — Master's student, National Research University Higher School of Economics, Moscow, Russia; Head of Scoring Systems Department Bank "Renaissance Credit", Moscow, Russia
https://orcid.org/0000-0002-1836-1555
aasmirnova_24@edu.hse.ru

***Authors' declared contribution*:**
S.V. Afanasiev — problem statement, conceptualisation of the article, critical analysis of the literature, description of the results and formation of the conclusions of the study.
D.M. Kotereva — development of methodology, program code, statistical data collection, tabular and graphical presentation of results.
A.A. Mironenkov — description of the results and formation of the conclusions of the study.
A.A. Smirnova — critical analysis of the literature, development of the program code.

Conflicts of Interest Statement: The authors have no conflicts of interest to declare.

The article was submitted on 11.02.2022; revised on 25.02.2022 and accepted for publication on 27.12.2022. The authors read and approved the final version of the manuscript.