**Investigation of Algorithms for Converting Dimension of Feature Space in Retail Data Analysis Problems**

Nikita V. Popov1, Natalya V. Razmochaeva2, Dmitry M. Klionskiy3,

Department of Software Engineering and Computer Applications

Saint Petersburg Electrotechnical University "LETI"

St. Petersburg, Russia

1nvpopov@stud.eltech.ru, 2nvrazmochaeva@etu.ru, 3dmklionsky@etu.ru

*Abstract* — In this paper, the goal was to investigate and determine the most accurate and fast algorithms for reducing the dimension of the feature space in the task of analyzing retail data, because large databases may contain noisy or duplicate information that should be eliminated to improve the quality of data processing. To do this, a review of existing solutions for selection and extract features was conducted, and strengths and old sides were highlighted. The results of this work have shown that the most accurate and fast algorithms are one-dimensional selection, which uses Chi-square as a static criterion, and the method of the scikit-learn SelectFromModel library, which accepts estimates of the parameters of the logistic regression model.

Keywords: data selection, client data analysis, dimension of feature space, machine learning

# **Introduction**

Today, data describing a particular subject area often contains a large number of different features that determine the properties of any processes or items. At the same time, these data sets can reach enormous sizes, which makes working with data very resource-intensive and time-consuming processes. Reducing the dimension of the feature space entails decrease memory used and getting rid of noisy and duplicate information. Various machine learning algorithms are used to achieve this goal, each of which has its own individual degree of efficiency and validity for its application area. Retail databases are no exception, they often contain large amounts of information, some of which are redundant. Therefore, determining how to reduce the feature space for more accurate analysis of retail data is a problem that only becomes more relevant over time. The object of the study is algorithms for reducing the dimension of the feature space, and the subject is various characteristics of the algorithms, such as the speed of operation, accuracy, speed, and visibility of the result. The article solves the following problems:

* Review of existing feature space reduction algorithms.
* Formation of criteria for determining optimal algorithms.
* Review of the algorithms considered in the article.
* Study of the characteristics of the considered algorithms.
* Comparison of the effectiveness of algorithms according to the identified criteria.

# **Comparison of analogies**

In order to determine the best algorithms, it is necessary to review modern solutions for reducing the number of parameters. Attention was focused on work in all areas of activity related to features, including medicine, equipment development, etc. Theoretical calculations and practical solutions were considered.

## Scientific article "Methods for reducing the dimension of statistical data space" [1].

Several ways of reducing the dimension of features are considered:

* Principal component analysis.
* Factor analysis.
* Multidimensional scaling.
* Algorithms for reducing dimension in automated system-cognitive analysis.

The theoretical description of algorithms and methods of their estimation is given. The algorithms considered extract features, but not reduce them, so their use in working with retail data is questionable.

## Scientific article "Reduction the dimension of the feature space in problems of identifying radiating objects using radio monitoring data using artificial neural networks" [2].

Principal component analysis and the use of autoassociative neural networks was examined. Analysis of the capabilities of artificial neural networks has shown that autoassociative neural networks can minimize the feature space without significant loss of information in working with radio signals.

## Scientific article "Algorithms and methods for reduction space of diagnostic features" [3].

A comparative analysis of methods for reduction dimension of the space of diagnostic signs in medical practice, such as:

* Cluster analysis.
* Principal component analysis.
* Method of extreme grouping of paramenters.
* Multidimensional scaling.

The theoretical description of the methods is given. Despite the research of algorithms in the field of medicine in this article, these algorithms can be applied to other areas.

## The scientific article "Algorithm GRAD for selection of informative subspaces of features" [4].

The algorithms AlDel, Grad, and the combined DX algorithm for reduction feature space are analyzed. Experimental calculations were carried out to test the effectiveness of the methods. It is concluded that the GRAD algorithm based on the AdDel method is most appropriate.

## Scientific article "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning" [5].

The article describes widely used techniques for selection and extraction of features. Descriptions of methods for selecting features are given: Correlation Coefficient, BW-ration, PAV, mRmR, I-RELIEF, CMIM, INTERACT, Genetic Algorithm, SVM-REF.

Descriptions of methods for distinguishing features are given: Independent Component Analysis, Principle Component Analysis, Nonlinear Principle Components Analysis, Probabilistic Principle Component Analysis, Kernel Principle Component Analysis.

It was concluded that for tasks such as disease detection, methods for detecting features are more convenient, while reduction algorithms are designed to get rid of noisy and duplicate signs.

# **Comparison criteria for analogs**

* Ability to configure algorithms.

Algorithms can work on data of different sizes with different efficiency, so you need to be able to configure the algorithm parameters.

* The number of the considered algorithms.

A large number of algorithms allows you to determine as many effective strengths as possible for comparison.

* Type of reduction algorithm.

Algorithms can not only reduce the number of features, but also transform them into new ones, which may cause the data to lose its visibility for analysts. Therefore, you should distinguish between reduction and selection methods.

Comparison by criteria is presented in table. 1

TABLE I. Comparison by criteria

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № analogy | 1 | 2 | 3 | 4 | 5 |
| Ability to configure algorithms | + | + | + | - | + |
| Number of algorithms | 4 | 2 | 4 | 3 | 14 |
| Type of reduction | extr | extr | extr | select | select and extr |

The above articles describe in detail most of the modern methods of reduction dimension of the feature space, but it is also worth to noting that there is no practical implementation of the algorithms, so it was not possible to determine quantitative comparison criteria. Currently, algorithms have been developed for specific areas, such as medicine or research of physical phenomena, but these applications do not cover the entire area of solving the problem of a large number of parameters in various fields of production and science, including retail databases.

The algorithms of first, second, and third articles are methods for extraction features, so their use in analyzing retail data may not correspond conditions. Our goal is to determine the characteristics of the algorithm that works for client data, so the second and third articles based on the research of radio signals and medical dignoses are less relevant in our work.

Choosing a solution method

Based on the results of the removal of existing analogues, requirements were formed for one method or several methods that could bring customer data in a form suitable for further analysis - that is, to reduce all noisy and duplicating features. At the same time, the developed methods should not depend on the number of features or the size of the analyzed data, but should be able to configure the algorithm parameters for specific data. Algorithms should have high accuracy of results, high speed of operation, and partially borrowed strengths of existing reduction algorithms. In this case, specifically for the task of analyzing client data, the focus should be on methods for select features, rather than extract them.

Description of the solution method

Description of the considered algorithms for reducing the feature space.

To reduce the number of features, we use methods such as reducing features with low variability, univariate feature selection, and the principal component analysis. Methods for selecting features such as the SelectFromModel function of the scikit-learn library and the recursive feature elimination algorithm are also considered. These methods use the results of the estimator models presented in this paper by logistic regression, random forest, and decision extra trees.

## Reduction of features with low variability.

All objects whose variance does not correspond a certain threshold are deleted. By default, the method deletes all objects with zero variance, i.e. objects that have the same value in all samples.

## Univariate feature selection.

This method selects the features that have the most pronounced relationship with the target variable using the Chi-square statistical criterion. The Chi-square formula is derived in equation (1):

|  |  |
| --- | --- |
|  | (1) |

## Principal component analysis (PCA).

The method reduces the dimension of the data feature space by using a linear algebra-based transformation by projecting data into a smaller dimension space. The calculation of the main components is reduced to the calculation of eigenvectors and eigenvalues of the covariance matrix of the source data. This algorithm is the only one among the presented algorithms for teaching without a teacher, i.e., it does not need a target parameter for its operation.

## SelectFromModel Method.

The method takes as an input model-evaluator, which provides values for each of the parameters that are compared with the parameter of the "threshold" function. Features that are below the threshold are removed from the data.

## Recursive feature elimination method (RFE).

The algorithm receives as an input evaluator model that provides values for each of the parameters. Based on these values, features are removed one by one from the original set. Considering all smaller sets of features, the algorithm saves only the most important ones until the number of features defined by the "n\_features\_to\_select" parameter remains in the set.

## Logistic regression.

Logistic regression is used to predict the probability of occurrence of an event based on the values of a set of features. To do this, a dependent variable that takes the values 0 or 1 is built and a set of independent variables which required to calculate the probability of accepting a particular value of the dependent variable.

In the logistic regression model, there are weights of features that can be considered as feature important in the problem of reduction the dimension of the space of these features. The closer the weight is to zero, the less influence this parameter has on the decision.

## Random forest.

A random forest is a model consisting of many decision trees. The decision tree is a decision support tool. The structure of the tree is "leaves" and "branches". The" branches "of the solution tree contain attributes that the target function depends on, the" leaves " contain the values of the target function, and the other nodes contain attributes that differ in cases. Classification is performed by descending the tree to the leaf and issuing the corresponding value.

In each node, the decision tree looks for the value of a certain parameter that will lead to the maximum reduction of Gini impurity. Gini impurity - the probability of incorrect marking in the node of a randomly selected sample. The Gini impurity of a node equal 1 minus the sum of the class's relations to the total number of samples squared for each of the set of classes.

A random forest is a model in which a random set of samples from a data set is selected to build a set of trees. While splitting nodes, random sets of parameters are selected. The significance of a parameter in a random forest is the total reduction of Gini impurity in all nodes that use this parameter for partitioning. This value is the most important attribute that can be used to judge whether it is needed in the data set.

## Decision extra trees.

In the decision extra trees, as in random forests, a random subset of possible features is used, but instead of searching for the most optimal thresholds, threshold values are randomly selected for each possible feature, and the best of these randomly generated thresholds is selected as the best rule for node separation.

# **Research of characteristics of considered algorithms**

To compare the methods described above, measurements of the speed and accuracy of these algorithms were performed. Python was chosen as the programming language because it has developed a large number of libraries that are widely used in machine learning, including the scikit-learn library, which contains some of the data selection algorithms.

The data set [6] was used as data, containing 17 features and 5300 tuples after the initial elimination of incomplete information. Since all the presented algorithms, except for the reduction of features with low variability and PCA, are algorithms for teaching with a teacher, the feature defined as the target was selected.

Each set of data obtained as the result of one of the algorithms is used to build a model based on logical regression. Next, the target parameter of the test data is calculated (20% of the total number of the set) and compared with the original one. The accuracy of the feature reduction algorithm is determined by the ratio of correct predictions to all.

## Reduction of features with low variability.

In the scikit-learn library, the method is implemented by the sklearn.feature\_selection.VarianceThreshold (version v0. 22) [7]. The results of applying the method on the used data set are presented in table. 2. The first result in table. 2 characterizes the forecast without using this method.

TABLE II. Results of applying the VarianceThreshold method

|  |  |  |  |
| --- | --- | --- | --- |
| Threshold | Number of excluded features | Accuracy | Operating time |
| – | – | 0.809606 | – |
| 0 | 0 | 0.809606 | 0.003509 |
| 0.005 | 1 | 0.808733 | 0.003040 |

As the results showed, the accuracy of the model only worsened with use of the method.

## Univariate feature selection.

Function sklearn.feature\_selection.SelectKBest (version v0. 22) [7] was used to implement the method with parameter score\_func = chi2. The accuracy for different data sets after performing the reduction using this algorithm is shown in Fig. 1. The operating time of the method is 0.006986. At the highest points (with the number of features 11, 12, and 14), the curve has a value of 0.810480.

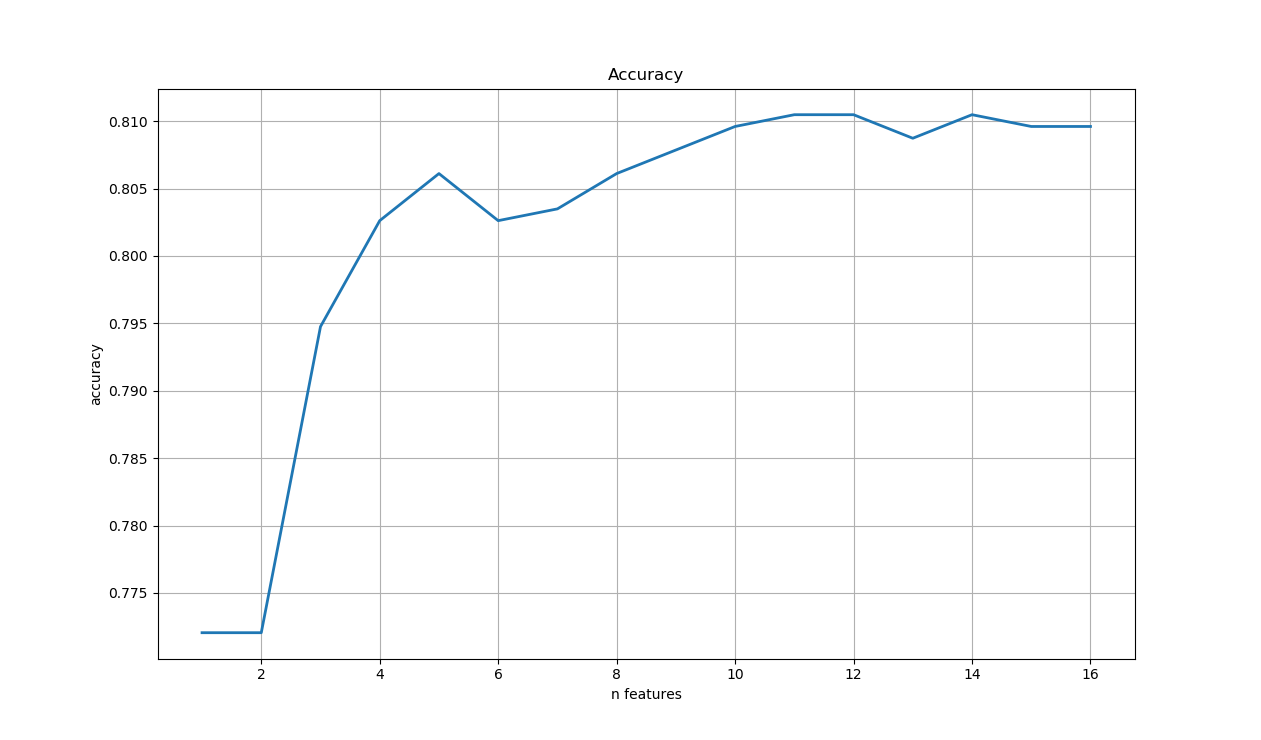


Figure 1 - Accuracy of univariate feature selection for a different number of features

## Principal component analysis.

Method is implemented by the sklearn.decomposition.PCA (version v0. 22) [7] in the scikit-learn library. Recorded results are shown in Fig. 2. The speed of the algorithm is 0.011030 sec.

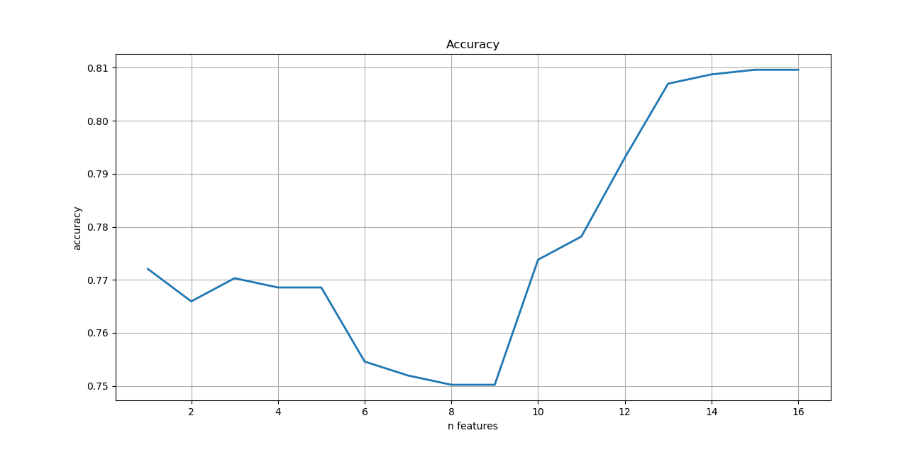


Figure 2 - Accuracy of the PCA for a different number of features

## SelectFromModel method and recursive feature elimination.

In the scikit-learn library algorithms has an implementations by the sklearn.feature\_selection.SelectFromModel (version v0. 22) [7] and sklearn.feature\_selection.RFE (version v0. 22) [7] respectively.

Models of logistic regression, random forest classifier, and decision extra trees are represented by sklearn.linear\_model.LogisticRegression (version v0. 22) [7], also known as LR, sklearn.ensemble.RandomForestClassifier (version v0. 22) [7], also known as RFC, and sklearn.ensemble.ExtraTreesClassifier (version v0. 22) [7], also known as ETC, respectively. The results of the algorithms efficiency are presented in the table. 3. and fig. 3-4.

TABLE III. Results of the SelectFromModel and RFE methods

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | SelectFromModel | | | RFE | | |
| Model | LR | RFC | ETC | LR | RFC | ETC |
| Speed | 0.057502 | 0.237182 | 0.167122 | 0.487762 | 3.387273 | 2.251238 |
| Accuracy | 0.810480 | 0.809606 | 0.809606 | 0.810480 | 0.809606 | 0.809606 |

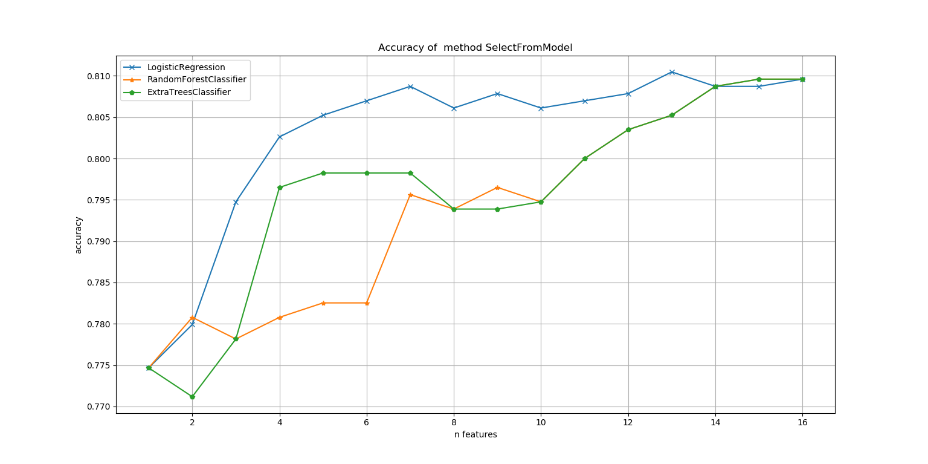


Figure 3 - Accuracy of the SelectFromModel method for a different number of features

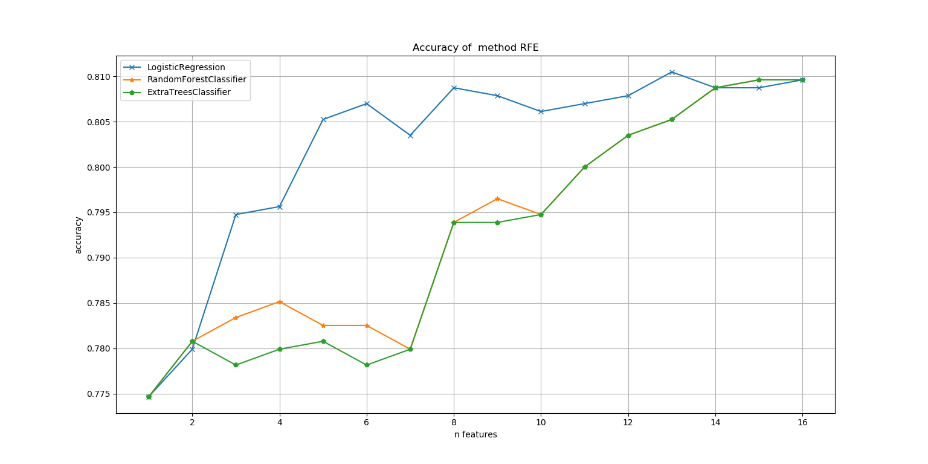


Figure 4 - Accuracy of the RFE method for a different number of features

The logistic regression and random forest classifier models for the SelectFromModel and RFE are almost identical, so to compare these algorithms, a graph based on the decision extra trees model was drawn on the Fig. 5.

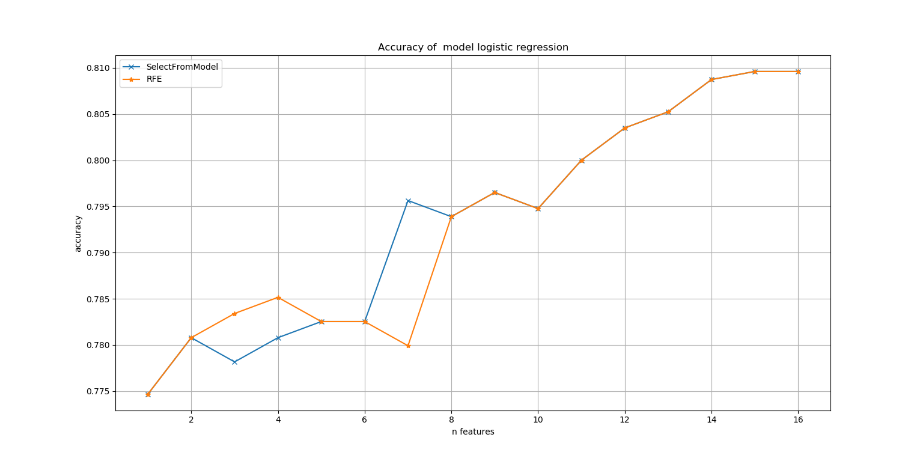


Figure 5 - Accuracy of the decision extra trees model for a different number of features

# **Comparison of effectiveness of the algorithms.**

Applying of the feature reduction method with low variability led to a decrease in the accuracy of the solution. This was due to the high variability of the used data. Therefore, this method is more effective for larger data that has several dozen features.

Univariate feature selection showed the highest level of accuracy among the considered algorithms – 0.810480.

The principal component analysis showed high accuracy only when the feature space is reduced by 1-3 features, while the accuracy decreases with further reduction.

In the SelectFromModel and recursive feature elimination algorithms, logistic regression showed the highest accuracy among models with a slight margin. For the first method, the random forest model performed the worst, and for the second decision extra trees.

If we compare these two algorithms, the results are almost identical, except for the case of using the decision extra trees classifier, where the average accuracy of the SelectFromModel method was higher (by 0.6%)

The algorithms for reducing features with low variability, univariate feature selection, principal component analysis and SelectFromModel, which implements the estimation of logistic regression, showed a shorter working time-less than 0.1 seconds. RFE showed the longest operating time, which uses a random forest model.

# **Conclusions**

In the course of the work, it was found that the greatest accuracy (0.810480) of reducing the dimension of the feature space has a univariate feature selection that uses the Chi-square as a static criterion. The speed of this algorithm is high compared to other considered algorithms. Less accurate results were shown by recursive feature elimination and SelectFromModel methods that use logistic regression, random forest, and super-random tree estimators. For the first and second models, the average accuracy of the algorithms was approximately the same, but for decision extra trees the SelectFromModel method was 0.6% more accurate. Also, logistic regression showed the highest accuracy among the estimator models (1.2% more than random forest and 1% more decision extra trees). The speed of these algorithms relatively to others was low, but it can be regulating by the number of evaluators in the models. The reduction of features with low variability and the PCA showed high speed, but low accuracy. Except for it, all considered algorithms provide the ability to adjust the number of saved parameters, and number of estimators for models for adjustment the accuracy and speed of the algorithms.

As a promising vector of development, we can determine research of other modern algorithms for preprocessing and data analysis and improvement of the methods considered in the article: combining methods or using neural networks.

##### References

1. Orlov Alexander Ivanovich, Lutsenko Yevgeny Veniaminovich methods for reducing the dimension of statistical data space / / Scientific-scientific journal of Kubgau. 2016. No. 119. URL: https://cyberleninka.ru/article/n/metody-snizheniya-razmernosti-prostranstva-statisticheskih-dannyh (accessed 23.12.2019).
2. With Adzhemov. S., Tereshonok M. V., Chirov D. S. Reducing the dimension of the characteristic space in the identification of radiating objects using radio monitoring data using artificial neural networks / / T-comm. 2008. No. 6. URL: https://cyberleninka.ru/article/n/snizhenie-razmernosti-priznakovogo-prostranstva-v-zadachah-identifikatsii-izluchayuschih-obektov-po-dannym-radiomonitoringa-s (accessed 23.12.2019).
3. Maksyuta N. V., and povoroznyk. I. Algoritmi and methods for reducing the space of diagnostic signs // Vestnik NTU KhPI. 2005. No. 46. URL: https://cyberleninka.ru/article/n/lgoritmy-i-metody-snizheniya-prostranstva-diagnosticheskih-priznakov (accessed 23.12.2019).
4. Zagoruiko N. G., Kutnenko O. A. Url Grad for selecting an informative feature subspace / / url: http://math. NSC. ru / ~ wwwzag / 7. HAIL. pdf (date accessed: 12.01. 2016). - 2005.
5. Khalid S., Khalil T., Nasrin S. review of methods of feature selection and feature extraction in machine learning / / scientific information conference 2014. - IEEE, 2014. - Yeah. 372-378.
6. Bank marketing // URL: https://www.kaggle.com/henriqueyamahata/ bank-marketing-classification-roc-f1-recall/data (accessed 23.12.2019).
7. Library scikit-learn // URL: https://scikit-learn.org (accessed 23.12.2019).