A. V. Goncharov[1], **V. V. Strijov**[2]

Interpreted deep learning models in the social ranking tas*

The task of credit risk modeling is one of the most important parts and main activities of financial institutions. Person or company gets the decision about loan providing based on the model results. This paper solves the tasks of scoring card construction and credit scoring model building. Intra-bank analytics and regulators requirements leads to the necessary of using the interpreted classification models. In particular, the logistic regression model. Legal purity and claims from clients cause this restriction. The purpose of this paper is feature space transformation and expanding for quality improving with a minimal complexity increase. The point is constructing the solution as a deep learning superposition that allows including all the necessary feature processing procedures. The main paper idea is constructing a single optimization procedure for sequential superposition using interpreted procedures of feature processing and model constructing. The segmentation, grouping and feature generation procedures are combained in one superposition. The paper proposed to replace the local quality criteria with a global one in the iterative optimization procedure. The optimization procedure uses the AUC quality criterion for achieving the best solution. The quality of the proposed solution is evaluated with the open credit scoring data and compared with the base model and complex noninterpreted models.

Key words: social ranking; scoring card; credit scoring; deep learning; models superposition; feature generation; parameters optimization.

1 Introduction

The credit scoring task solving is the basis activity for financial institutions, consumer lending banks and retailers. The solution is a risk assessment model for a borrower or existing loan portfolio. The purpose is making a decision of loan granting based on default estimation. In the model construsting task the borrower's probability of default is determined by his personal information.

The analytical departments are working on scorecards construction. Scorecards contain information for making a decision. Since their using is the business basis, their development is the key competence of retail bank risk management.

[1]Moscow Institute of Physics and Technology, master., E–mail: alex.goncharov@phystech.edu

[2]Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Russia; E–mail: strijov@ccas.ru

*Работа выполнена при финансовой поддержке РФФИ, проект № 16-07-01158

The analyst is not able to identify manually a large number of data patterns. Such a complex task uses the automatisation technologies for the clients assessment process, combining different scoring and risk models.

The model performance is very important. Improving the forecast accuracy can lead to significant resources saving. Therefore, banks need the analysis of various scorecards development models and algorithms.

Bank managers appreciate the validity of the scoring score because they avoid scandalous situations around the bank. This leads to the interpreted models using inside the banking structures. Such models show a lower quality than their uninterpreted analogues based on neural networks or uninterpreted model ensembles. The governing bodies implement external constraints [6] on models constructing. They must satisfy the dependencies in the scoring cards, which leads to the requirement of their interpretability.

There are two approaches of scoring models construction. The modern one uses new, uninterpreted technologies that allow to obtain high quality [8, 9, 10, 11]: neural networks, boosting on decision trees. Analysts use the listed models to confirm or disprove the interpreted results. While the results inconsistency the loan request is sent for double-check analysis.

A common one approach is based on the interpretable classifying models, such as logistic regression [12, 13, 14, 15, 16]. The literature proposes methods for feature spaces generation. The resulting models allow us to identify the necessary interpreted features for approving by the bank analysts.

Scorecard is used to determine borrowers creditworthiness. It consists of linear, categorical and binary features. Analysts conduct social studies to determine the impact of various features. They include the most common features that are simply to verify. Analysts perform several procedures while model construction. The subsequent processing includes all of them: the segmentation and grouping procedures [17], interpretable feature combinations creating, the classification models construction and the quality criterion calculation. The basic principle of model constructing is a superposition of these procedures.

The [4] paper suggests to use segmentation for feature space processing. Segmentation maps a linear feature into categories. Analysts consider this step as a stable and interpretable solution for the hidden nonlinearities problem [5]. Using methods of [4] present paper creates a basis for the segmentation procedure. In [2] the metric classification model are suggested for processing small data volume. The initial data clustering improves the quality without interpretability loss. Such approach allows to solve the problem of the heterogeneities presence in the data. This generalizes results of [7]. The present paper carries out a study on the feature space transformation and do not uses object space separation methods.

The study is relevant for the financial organizations needs of scoring models construction,

quality improving without complexity increase and interpretability loss. There is no single procedure for scoring model constructing. It makes possible to unite into a single superposition several feature generation proocedures.

There are two types of quality criteria in the credit scoring tasks: optimization and operational ones. The industry set the operational quality criteria while solving business tasks. Optimization criteria affect the model parameters configuration and are necessary for its construction, but they are not related to operational ones.

Operational quality criteria:

- AUC — is the area under the ROC curve (4). It is a quality criterion in the binary classification task that has a physical sense: the fraction of correctly ordered pairs of objects. This criterion is used for model optimization and its expert evaluation. It is the main comparing criterion.

- Structural complexity, the number of features. The model interpretability directly depends on its complexity. Experts define the relevant requirements for model interpretability. In the proposed model the criterion depends on the structural parameters value, therefore thus criterion (4) is not included in the optimized functional.

- Non-negativity of the model features weights. The requirement of non-negativity of weights allows to analyze the used features.

- Stability in time. Such a restriction is caused by high risks while unstable model behavior.

Optimization quality criteria:

- Likelihood of the model. The likelihood indicates the probabilistic meaning of the obtained answer and is used in solution constructing for the optimal logistic regression parameters search.

- Stability — the features orthogonality. The features multicollinearity is necessary for the model stability. The more heterogeneous features are, the better the model behave is.

- Precision, recall. The method precision and recall let calculate the financial risks of the model in some business applications.

Paper proposes using the operational quality criterion in a single optimization problem. It studies the applicability of the proposed solution and compares with the basic and alternative

approaches. The [3] reviews the comparison methods and proposes a comprehensive methodology for their estimation. AUC or area under the ROC curve is the most common quality criterion for evaluating scoring models. In paper [1] it is proposed to use AUC as the target and the only quality criterion in the optimization problem. The same approach is used in this paper.

The aim of the paper is constructing a deep learning model for scorecard formation as the interpreted feature generation procedure superposition. This model uses the deep learning principle: optimizing the quality criterion within each individual procedure. It is necessary to satisfy the requirements of interpretability and final model quality for a given low complexity. This study proposes an iterative algorithm for the entire superposition optimization by modifying the segmentation, grouping, and feature generating procedures.

2 The social ranking task statement

Let $D = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, m\}$, be a set of m objects. Linear, categorical and binary features describe each of them $\boldsymbol{x}_i \in \mathbb{R}^l \times \mathbb{C}^c \times \mathbb{B}^b$. There are two possible classes for each object: $y_i \in \{0, 1\}$. Object indexes $\{i = 1, \ldots, m\} = \mathcal{I}$ are divided $\mathcal{I} = \mathcal{L} \bigsqcup \mathcal{T}$ into the learning and testing indexes.

By $\boldsymbol{X} \in \mathbb{R}^{m \times n}, \boldsymbol{X} = (\boldsymbol{\chi}^1, \ldots, \boldsymbol{\chi}^n)$ denote a ??PLAN MATRIX??, where χ_j — feature j. Let $\mathcal{A} = \mathcal{A}_l \sqcup \mathcal{A}_c \sqcup \mathcal{A}_b$ be a feature indexes set, where $\mathcal{A}_l$ denotes linear features indexes, $\mathcal{A}_c$ — categorical features, a $\mathcal{A}_b$ — binary ones.

The classification model F is presented in the superposition form $F = f \circ f_f \circ f_g \circ f_s$, where f_s is a segmentation procedure for linear features, f_g — grouping procedure for categorical features, f_f constructs new features based on the initial ones. All these functions translate object $\boldsymbol{x}$ from initial space to the new one. This causes a change in the ??PLAN MATRIX?? $\boldsymbol{X}$. Superposition $f_f \circ f_g \circ f_s$ translates one ??PLAN MATRIX?? to another:

$$f_f \circ f_g \circ f_s : \boldsymbol{X} \mapsto \hat{\boldsymbol{X}}, \tag{1}$$

where $\hat{\boldsymbol{X}} = (\hat{\boldsymbol{\chi}}^1, \ldots, \hat{\boldsymbol{\chi}}^{\hat{n}})$.

The paper uses logistic regression model as f function because of its interpretability over feature space:

$$f(\hat{\boldsymbol{X}}, \boldsymbol{w}) = \frac{1}{1 + \exp\left(-\hat{\boldsymbol{X}}\boldsymbol{w}\right)}. \tag{2}$$

The main optimization criteria $L(\boldsymbol{w})$ for logistic regression model is logarithmic likelihoog function:

$$L(\boldsymbol{w}) = -\ln P(D|\boldsymbol{w}) = -\sum_{i \in \mathcal{L}} \big(y_i \ln \boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + (1 - y_i) \ln(1 - \boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i)\big). \tag{3}$$

The area under ROC curve is the criteria for segmentation, grouping and feature generation procedures optimization and quality control due to the bank requirements:

$$Q(w) = \frac{\sum_{i\in\mathcal{L}}\sum_{j\in\mathcal{L}}[y_i < y_j][F(\boldsymbol{x}_i) < F(\boldsymbol{x}_j)]}{\sum_{i\in\mathcal{L}}\sum_{j\in\mathcal{L}}[y_i < y_j]}. \tag{4}$$

Function f is parametric with $\boldsymbol{w}$ parameters. Functions f_f, f_g and f_s include parameters $P = p_f \sqcup p_g \sqcup p_s$ and structural parameters $H = h_f \sqcup h_g \sqcup h_s$. The optimal parameters $\boldsymbol{w}, P$ and structural parameters H selection task for classification model F is an optimization task:

$$H^*, P^*, w^* = \underset{H,P,\boldsymbol{w}}{\operatorname{argmin}} Q(\boldsymbol{w}). \tag{5}$$

3 Parameters and model structure optimization task

The baseline method for solving 5 task is to optimize local quality criteria for each procedure and depicted as the structural diagram in the Figures 1 and 2:

$$h_s^*, p_s^* = \underset{h_s,p_s}{\operatorname{argmin}} S_s(h_s, p_s), \tag{6}$$

$$h_g^*, p_g^* = \underset{h_g,p_g}{\operatorname{argmin}} S_g(h_g, p_g), \tag{7}$$

$$h_f^*, p_f^* = \underset{h_f,p_f}{\operatorname{argmin}} S_f(h_f, p_f), \tag{8}$$

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmin}} L(\boldsymbol{w}), \tag{9}$$

where S_s, S_g, S_f are local quality criteria for baseline method and are shown below.

The model F learning process is following. Learning set is given. The method consistently selects parameters for each function from superposition $F = f \circ f_f \circ f_g \circ f_s$, that are optimal for the corresponding quality criterion. The selection procedures are independent.

A significant change in some function parameters does not lead to a change in other functions parameters. Such independence in parameters space and local quality criteria optimization that are not related to the operational criterion lead to the solution nonoptimality.

Statement 1. For optimization task: $S(x_1, \ldots, x_n) \to \min$ *with initial point* $X^0 = x_1^0, \ldots, x_n^0$ *the optimization procedure:*

$$\hat{x}_i = \underset{x_i}{\operatorname{argmin}} S_i(\hat{x}_1, \ldots, \hat{x}_{i-1}, x_i, x_{i+1}^0, \ldots, x_n^0), i = 1, \ldots, n,$$

delivers the quality not better, than optimization procedure:

$$\widetilde{x}_1, \ldots, \widetilde{x}_n = \underset{x_1,\ldots,x_n}{\operatorname{argmin}} S(x_1, \ldots, x_n),$$

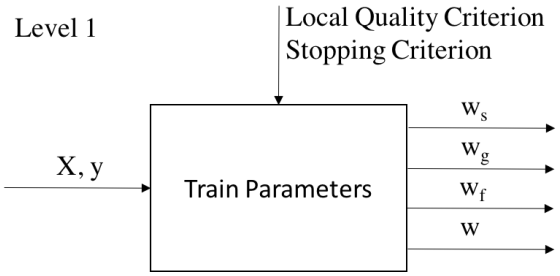

Рис. 1: Baseline method structural diagram, upper level

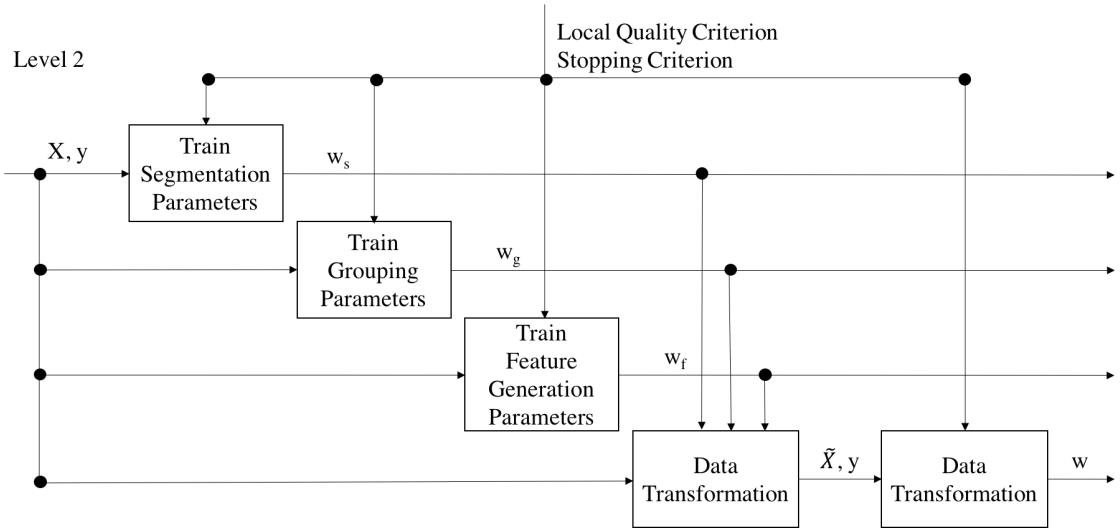

Рис. 2: Baseline method structural diagram, lower level

that is: $S(\hat{x}_1, \ldots, \hat{x}_n) \leq S(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

Proof: $\forall x_i \neq \widetilde{x}_i : S(\widetilde{x}_1, \ldots, x_i, \ldots, \widetilde{x}_n) \leq S(\widetilde{x}_1, \ldots, \widetilde{x}_i, \ldots, \widetilde{x}_n)$. $\square$

The optimization is a simultaneously procedure over the entire parameter space. It finds the optimal parameters for 5. This paper proposes to use a single criterion instead of local ones. It will be optimized consistently for each individual procedure. The structural diagrams in Figures 3 and 5 show this training procedure. The change in the baseline method is motivated by better solution search.

Each of the optimized variables $H, P, \boldsymbol{w}$ affects the final quality criterion while solving the problem (5). The optimization procedure in the basic approach is carried out sequentially for each group of variables. Paper proposes to create an iterative superposition optimization algorithm, where the initial model parameters approximation for the next iteration is the result of the previous one. Figures 4 and 5 show the structural diagram of the proposed algorithm.

Such a choice is assumed to be optimal: let $\hat{x}_1 \sqcup \cdots \sqcup \hat{x}_n = x_1^0 \sqcup \cdots \sqcup x_n^0$. The pseudocode $A(N)$, where N is the number of algorithm iterations, describes the task solving algorithm, где N — число итераций алгоритма:

$for \quad i = 1, \ldots, N :$

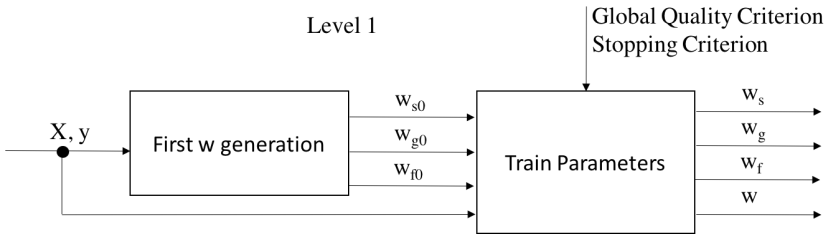

Рис. 3: Structural diagram of method with one quality criterion, upper level

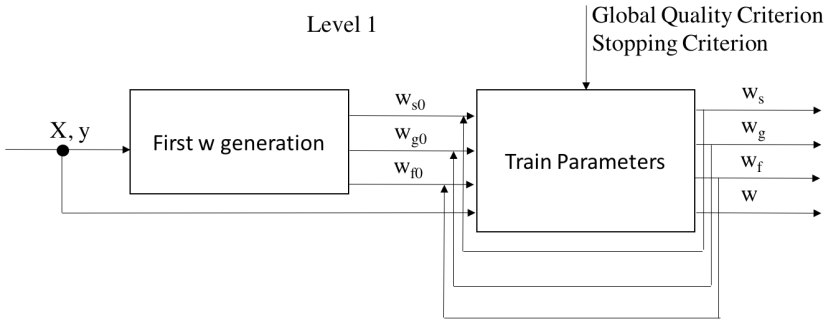

Рис. 4: Proposed method structural diagram, upper level

$$\hat{x}_1 = \underset{x_1}{\operatorname{argmin}}\,\text{step} \quad Q(\hat{x}_1 \sqcup \cdots \sqcup \hat{x}_n)$$

$\cdots$

$$for \quad i = 1, \ldots, N:$$

$$\hat{x}_n = \underset{x_n}{\operatorname{argmin}}\,\text{step} \quad Q(\hat{x}_1 \sqcup \cdots \sqcup \hat{x}_n).$$

The numerical optimization procedure $A(N)$ for solving the task $Q(x_1, \ldots, x_n) \to \min$, is not better, than: $for\ \ j = 1, \ldots, M:\ A(\frac{N}{M})$.

Matrices $\hat{\boldsymbol{X}}$ can be different on each cycle. This leads to selection ambiguity of the initial function f parameters. Proposed to perform an f optimization procedure in each cycle.

4 Linear feature segmentation

Segmentation allows to take into account the complex piecewise nonmonotonic dependence between the risk level and linear feature and to simplify the model interpretation. Initial feature space is supplemented by binary features that denote the feature belonging to a given segment on a linear scale. A non-linear dependence example is shown in Figure 6.

4.1 The segmentation problem statement

The segmentation procedure structural parameter h_s is a set of feature partition nodes number. Let feature χ belongs to an interval $[a, b]$. The segmentation procedure parameters p_s are the nodes coordinates set, and $p_s^j \in [a, b], j = 0, \ldots, h_s; \quad a = p_s^0 < \cdots < p_s^j < \cdots <$

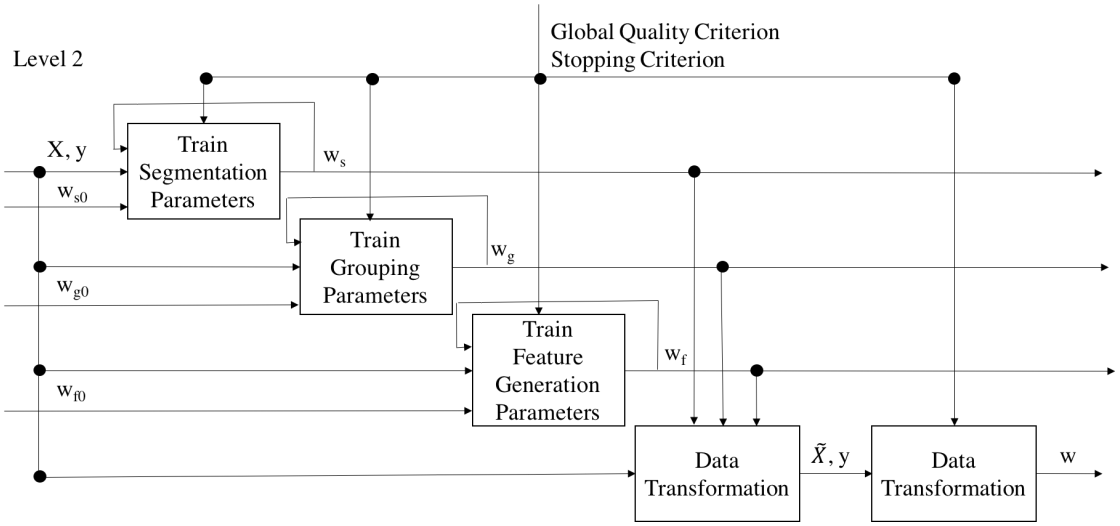

Рис. 5: Structural diagram of method with one quality criterion, lower level

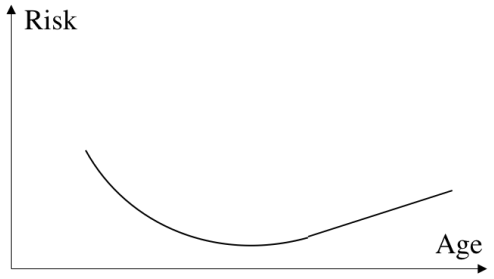

Рис. 6: A non-linear dependence example.

$p_s^{h_s} = b$.

The initial ??PLANmatrix?? X is updated with a new binary features set that are indicators of the feature value belongings for each segment:

$$\{\chi^j\}_{j=1}^{h_s} \in \mathbb{B}^{h_s} : \chi^j = \left[\chi \in [p_s^{q-1}; p_s^q]\right], q = 1, ..., h_s.$$

The segmentation procedure structure parameters and parameters selection on the learning set D occurs by solving the following optimization task:

$$h_s^*, p_s^*, w^* = \underset{h_s, p_s, w}{\operatorname{argmin}} Q(h_s, p_s, w | D, \mathcal{A}_l, \mathcal{L}),$$

where $Q$is a quality criterion (4).

4.2 Segmentation task solution

The partition nodes belonging to the linear scale are the segmentation parameters. The chosen segment q for feature i [**?**] WOE and set fraction are used for structural parameters and parameters selection:

$$\text{count}_q = \frac{1}{m} \sum_{j \in \mathcal{L}} \left[\boldsymbol{\chi}_j \in [p_s^{q-1}; p_s^q]\right],$$

$$\text{woe}_q = \log \frac{g_j}{b_j} = \log \frac{\sum_{j \in \mathcal{L}} [\boldsymbol{\chi}_j \in [p_s^{q-1}; p_s^q]] [y_j = 1]}{\sum_{j \in \mathcal{L}} [\boldsymbol{\chi}_j \in [p_s^{q-1}; p_s^q]] [y_j = 0]},$$

where g_j and b_j are fractions of 0 and 1 classes in the j segment.

For baseline solution (6) similar to [4] proposed to divide interpreted feature into maximum possible number of segments. Then combine them due to several principles: the objects portion is not less than 4% of the total number, WOE value is not significantly different for the neighboring segments. Thus the local quality criterion S_s for feature χ: $S_s = [count_q \geq 0.04] \quad [|woe_q - woe_{q-1}| \geq 0.1]$.

The paper novelty is using numerical optimization for segmentation parameters search. The target critetion is $Q(w)$, where w denotes all the model parameters. Feature χ optimization parameters are p_s, therefore the target criterion is $Q(p_s)$.

Optimization iteration for feature χ segmentation parameters is a gradient descent in the model parameters space:

$$p_s^i = p_s^{i-1} - \lambda \bigtriangledown Q(p_s),$$

where $\bigtriangledown Q(p_s)$ is calculated with numerical methods.

Optimization iteration for model segmentation parameters is several cycles of successive gradient descent iterations for each of $\chi \in X$. Segmentation parameters correction procedure takes a place after each numerical optimization iteration

$$Correction(p_s): \quad \text{удаляется} \quad p_s^q \quad \text{, если} \quad p_s^q - p_s^{q-1} < 2 \tag{10}$$

Thus, the structural segmentation parameter for feature χ changes while segmentation parameters consolidation at one linear space point.

$for \quad iteration = 1, \ldots, N:$

$\quad for \quad \chi \in X:$

$\quad\quad p_s^i = p_s^{i-1} - \lambda \bigtriangledown Q(p_s)$

$\quad\quad p_s^i = Correction(p_s^i) \qquad (10)$

Definition. The feature χ segmentation procedure complexity is the number of nodes or the structural parameter h_s value.

Definition. The model segmentation procedure complexity is the number of nodes or sum of the structural parameters values $\sum_{\chi \in X} h_s$.

5 Categorical features grouping

There are categorical features with a large number of categories in scoring cards. Such unordered features processing increases the model complexity greatly. For example, the

categorical sign of a borrowers profession includes hundreds of categories. Such a number of processed binary features also promotes the models overfitting and reduces the interpretability of the constructed solutions. To solve the described problem this paper proposes to combine the categories into groups to reduce the model and feature space complexity without reducing the solution quality. This is a grouping procedure, it reduces the model complexity and increases its interpretability.

5.1 Grouping task statement

Let categorical feature χ has $|C|$ categories, where C is its set of categories. Grouping procedure structural parameter h_g of feature χ is the number of new categories groups, $|h_g| < |C|$. The grouping parameter is a surjective map $p_g : C \to h_g$. Initial feature space changes: categorical feature χ is replaced with categorical feature χ_h.

$$\begin{array}{lcccccccl} \chi & = & 1 & 2 & 3 & \dots & \mathrm{C} & & \mathrm{C} - \text{categories number} \\ & & \downarrow & \downarrow & \downarrow & & \downarrow & & \\ \chi_h & = & \gamma_1 & \gamma_2 & \gamma_3 & \dots & \gamma_C & |h_g| - \text{new categories number, } \gamma_i \in h_g \end{array}$$

The grouping procedure structure parameters and parameters selection on the learning set D occurs by solving the following optimization task:

$$h_g^*, p_g^*, w^* = \underset{h_g, p_g, w}{\operatorname{argmin}} Q(h_g, p_g, w | D, \mathcal{A}_c, \mathcal{L}),$$

where Q is the quality criterion (4).

5.2 Grouping task solving

The surjective maps are the grouping procedure parameters. WOE values and the sample fraction of the selected category c are used in searching the parameters for baseline method:

$$count_c = \frac{1}{m} \sum_{j \in \mathcal{L}} [\boldsymbol{\chi}_j = c],$$

$$woe_c = \log \frac{g_j}{b_j} = \log \frac{\sum_{j \in \mathcal{L}} [\boldsymbol{\chi}_j = c][y_j = 1]}{\sum_{j \in \mathcal{L}} [\boldsymbol{\chi}_j = c][y_j = 0]},$$

where g_j and b_j are the classes 0 and 1 portions in the c category. The baseline method proposes to combine categories where fraction is small with the closest in terms of WOE value.

To construct the basic solution (7) proposes to combine similar woe_c values into one group,as well as categories, where the objects fraction is less than 4%, combine with closest

woe_c value. So the local quality criterion S_g for categorical feature χ: $S_g = [count_c \geq 0.04] \prod_{i,j}[|woe_i - woe_j| \geq 0.1]$.

It is necessary to selet surjections that deliver the quality optimum. Numerical optimization methods are not suitable for solving such kind of problems. The paper novelty is the use of genetic algorithms principles to search the grouping parameters. The target criterion is $Q(w)$, where w denotes all the model parameters. The optimization parameters for feature χ are p_g, Therefore the target criterion is $Q(p_g)$.

Definition. Individual a_χ in the grouping procedure optimization task of feature χ is a sequence $\{\gamma_1, \ldots, \gamma_C\}, \gamma_i \in h_g$, which defines an arbitrary surjective map p_g.

Definition. The crossing procedure $crossing(a^1_\chi, a^2_\chi)$ for individuals a^1_χ and a^2_χ is a crossing of sequences: $\{\gamma^1_1, \ldots, \gamma^1_C\}$ и $\{\gamma^2_1, \ldots, \gamma^2_C\}$. The crossing of two sequences is defined by a random binary vector $\boldsymbol{bin} \in \mathbb{B}^{|C|}$:

$a^1_{12} = \{\boldsymbol{bin}_1\gamma^1_1 + (1 - \boldsymbol{bin}_1)\gamma^2_1, \ldots, \boldsymbol{bin}_C\gamma^1_C + (1 - \boldsymbol{bin}_C)\gamma^2_C\}$.

$a^2_{12} = \{\boldsymbol{bin}_1\gamma^2_1 + (1 - \boldsymbol{bin}_1)\gamma^1_1, \ldots, \boldsymbol{bin}_C\gamma^2_C + (1 - \boldsymbol{bin}_C)\gamma^1_C\}$.

The iteration of feature χ grouping optimization algorithm is a given number of random individuals crosses, where individuals are from the set $A = \{a^i_\chi\}$, and best individuals selection from the resulting set. The best individuals is the next approximation for p_g.

The iteration of the grouping in the model superposition optimization algorithm is several cycles of consecutive genetic algorithm iterations for each categorical feature $chi\ inX$. Since there is no restriction on the number of new categories for the initial initialization and individuals crossing, the grouping structural parameters will be found automatically.

$for \quad iteration = 1, \ldots, N:$

$\quad for \quad \chi \in X:$

$\quad\quad for \quad j = 1, \ldots, N:$

$\quad\quad\quad A \quad append \quad crossing(a^i_\chi, a^k_\chi)$, where i, k are random

$\quad\quad\quad A \quad = \quad select \quad M \quad best \quad of \quad A$

$\quad\quad\quad p_g \quad = \quad select \quad one \quad best \quad of \quad A$

Definition. The grouping procedure of feature χ complexity is the number of new categories $|h_g|$.

Definition. The grouping procedure complexity is the sum of new categories number $\sum_\chi |h_g|$.

6 Interpretability features construction

Complex models deliver the highest solution quality by revealing nonlinear data regularities and complex relationships between the features. Less complex interpreted models can take into account this kind of interconnection if new features are created on initial feature space basis. They are superpositions of initial features and interpreted operations.

6.1 Feature construction task statement

Let $\{f_i\}$ be a set of interpreted generative nonparametric functions over feature space. The table below demonstrates the relevant operations and their properties.

Describtion	Formula	in	N in	out
Negate binary	$\overline{x}$	bin	1	bin
Logarithm	$\log(x)$	lin	1	lin
Logistic sigmoid	$\frac{1}{1+\exp(x)}$	lin	1	lin
Square root	$\sqrt{x}$	lin	1	lin
Inverse	$\frac{1}{x}$	lin	1	lin
Multiplication	$x * y$	any	2	lin
Sum	$x + y$	any	2	lin

Let $\{\chi_i\}, i \in \mathcal{A}_l \sqcup \mathcal{A}_b$ be the linear and binary fatures. It is necessary to find the best combination of all admissible superpositions of features $\{\chi_i\}$ using the functions $\{f_i\}$.

Definition. Superposition is a representable as a tree formula, each node consists of function from $\{f_i\}$, each leaf — of feature from $\{\chi_i\}$. Superposition complexity is a number of using elements from$\{f_i\}$ and $\{\chi_i\}$.

The task is choosing the best combination $p_f = \{\mathbb{F}_i\}$ of superpositions from the set of all superpositions Σ. Structural parameters is the number of superpositions in $h_f = |\{\mathbb{F}_i\}|$ and maximal superposition complexity, parameters are the combination $p_f = \{\mathbb{F}_i\}$. The structural parameters and parameters selection on the training set D occurs by solving the following optimization problem:

$$h_f^*, p_f^*, w^* = \underset{h_f, p_f, w}{\operatorname{argmin}} Q(h_f, p_f, w | D, \mathcal{A}_l \sqcup \mathcal{A}_b, \mathcal{L}),$$

where Q is a quality criterion (4).

6.2 Feature construction task solving

The feature cinstruction procedure parameters are the superposition sets. The correlation coefficients between superposition and class value are used for selection in baseline model:

$$corr_g(p_f) = \sum_{\mathbb{F}_i \in p_f} (1 - corr(\mathbb{F}_i, \boldsymbol{y})^2).$$

The constructed features set quality is evaluated with $corr_g$ for baseline model (8). Thus, the local quality criterion S_f for the given task is: $S_f = corr_g$

It is necessary to choose a superpositions set that delivers a quality optimum. Numerical optimization methods are not suitable for solving such a problem. The paper novelty is the use of genetic algorithms to search for new superpositions. The target criterion is $Q(w)$, where w denotes all the model parameters. Optimization parameters are p_f, so the target criterion is $Q(p_f)$.

Definition. Individual a in the feature construction optimization task is a superpositions combination $\{\mathbb{F}_i\}$.

Definition. Crossing procedure $crossing(a_1, a_2)$ for individuals a^1 and a^2 is a crossing of superpositions combinations: $\{\mathbb{F}_i^1\}$ and $\{\mathbb{F}_i^2\}$. The crossing of two such sequences is given by a random binary vector $\boldsymbol{bin} \in \mathbb{B}^{|\{\mathbb{F}_i\}|}$:

$a_{12}^1 = \{\boldsymbol{bin}_i \mathbb{F}_i^1\} \cup \{(1 - \boldsymbol{bin}_i)\mathbb{F}_i^2\}$.

$a_{12}^2 = \{\boldsymbol{bin}_i \mathbb{F}_i^2\} \cup \{(1 - \boldsymbol{bin}_i)\mathbb{F}_i^1\}$.

The optimization algorithm iteration for feature construction task is several cycles of random individuals crossings $\{\mathbb{F}_i\}$ from the set $F = \{\{\mathbb{F}_i\}_j\}$. The best individual is selected as the next approximation of p_g:

$for \quad iteration = 1, \ldots, N:$

$F \quad append \quad crossing(a^i, a^k)$, где i, k — случайные

$F \quad = \quad choose \quad M \quad best \quad of \quad F$

$p_f \quad = \quad choose \quad one \quad best \quad of \quad F$

7 Algorithm of procedure superposition

The computational experiment considers two approaches to combining the above described procedures: the basic and the proposed. The basic approach involves the sequential execution of the basic segmentation procedure (6), the grouping (7) and the feature generation (8) and the achievement the local criteria optimum for each procedure. The described criteria are not directly related to the target criterion, which leads to a nonoptimality of the solution

constructed.

The section 5, 6 and 7 above describe the proposed optimization procedures. Section 4 describes their unification pronciple. The following steps are consistently performed:

1) starting points for gradient methods and initial generations for genetic algorithms initialization,

2) the gradient descent method steps for the segmentation procedure for each feature,

3) the genetic optimization iteration for the grouping procedure for each feature,

4) the genetic optimization iteration for the generation procedure,

5) return to step 2 while the quality converges.

A pseudo-code combining the corresponding procedures in a single algorithm is below:

$for \quad iteration = 1, \ldots, N:$

$for \quad \chi_i, \quad i \in \mathcal{A}_l:$

$p_s^i = p_s^{i-1} - \lambda \bigtriangledown Q(p_s)$

$p_s^i = Correction(p_s^i)$

$for \quad \chi_i, \quad i \in \mathcal{A}_c:$

$for \quad j = 1, \ldots, N:$

$A \quad append \quad crossing(a_\chi^i, a_\chi^k),$ where i, k are random

$A \quad = \quad select \quad M \quad best \quad of \quad A$

$p_g^i \quad = \quad select \quad one \quad best \quad of \quad A$

$F \quad append \quad crossing(a^i, a^k),$ where i, k are random

$F \quad = \quad select \quad M \quad best \quad of \quad F$

$p_f^i \quad = \quad select \quad one \quad best \quad of \quad F$

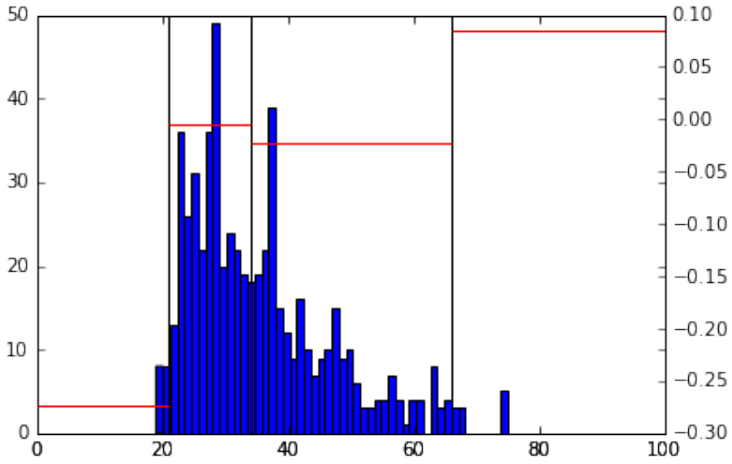

Рис. 7: The linear feature segmentation example.

8 Computantional experiment

The constructed optimization procedure as a single model structure is compared by a set of critteria with other models: neural networks, random forests, and boosting. The result of the computational experiment is the comparison of the models quality without the hyperparameters selection. The models are compared by AUC and the models structural complexity to assess their sustainability, interpretability and complexity. Interpretability of the final model is evaluated for the proposed method.

To conduct a computational experiment and demonstrate the interpretability of the proposed model results, the credit scoring dataset is selected from the UCI repository: German Credit Data. A subset of features was chosen. It contains: linear, categorical and binary features.

The segmentation optimization results for several linear features are shown in figures 7–9. These figures show the linear feature distribution in the histogram form, the proposed segmentation nodes and the weights of corresponding binary features. The linear features segmentation into groups is interpretable.

There are several conclusions from the segmentation modeling results. Borrowers can be separated into social groups due to the propensity to credit risk. For example, figure 8 shows that young and old people, from the bank perspective, are more risky.

Several experiments were conducted to evaluate the genetic algorithms convergence: the modeling results of the grouping procedure and the features generating procedure. There are examples below that demonstrate the built solution:

$$\begin{array}{lcccccccc} \chi & = & 1 & 2 & 3 & 4 & 5 & & 5 \text{ — the initial categories number} \\ & & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & & \\ \chi_h & = & 1 & 2 & 2 & 3 & 3 & & 3 \text{ — the new categories number, } \gamma_i \in h_{gi} \end{array}$$

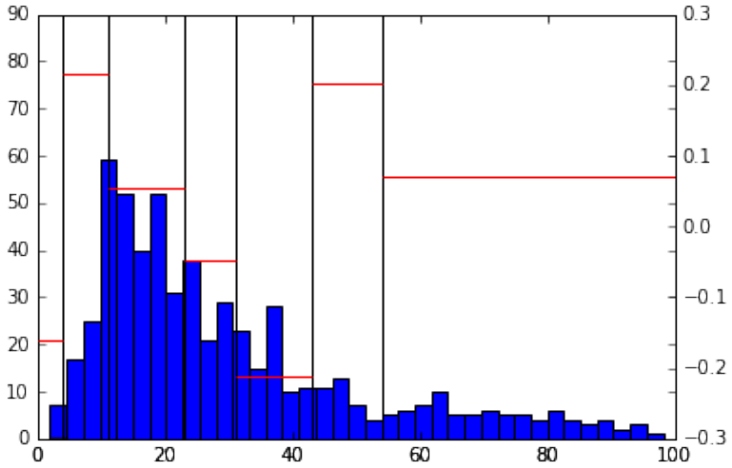

Рис. 8: The linear feature segmentation example.

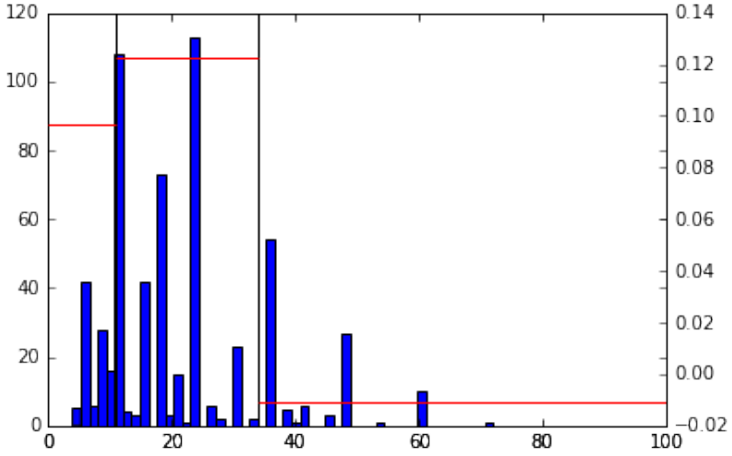

Рис. 9: The linear feature segmentation example.

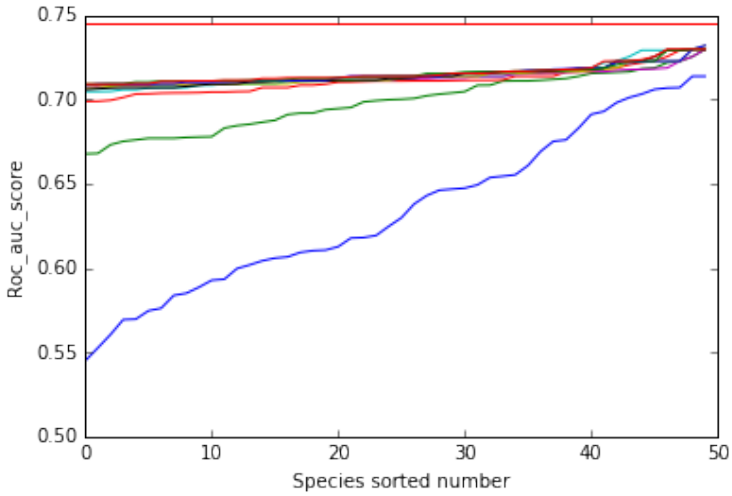

Рис. 10: The grouping procedure convergence.

$$\begin{array}{rcccccccl} \chi & = & 0 & 1 & 2 & 3 & 4 & 5 & 6 \text{ — the initial categories number} \\ & & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \\ \chi_h & = & 3 & 1 & 3 & 2 & 4 & 1 & 4 \text{ — the new categories number, } \gamma_i \in h_{gi} \end{array}$$

The model hyperparameters are selected automatically according to the described method. Generation includes individuals with a different number of categories. The optimal hyperparameters value delivers the best quality. The combination of categories leads to an increase in the models consistency, their interpretability and the quality. The proposed dataset does not include categorical features with a large number of categories, so the effect of reducing their number is negligible.

The figure 10 below shows an ordered by the quality generation. On the x-axis, the number of the individual in order of quality is laid down in one generation. The maximum value convergence to a stable value indicates the convergence and quality improvement of the final best individual. It is chosen as the optimal grouping parameter. The solution quality quickly converges to the optimal value, and the best individual determines the maximum quality.

We analyze the obtained partitions of the initial features. The partion with the maximum of the quality functional is interpretable. For example, a credit history feature containing 5 categories is broken down into three categories: "Past payment in the past" and "Critical account", "Did not take", "Pay out" and "Pay out due to this moment", which is an interpreted result: negative examples have merged into one category, and positive and neutral - into the other two.

A similar experiment was carried out for the feature generation procedure. Each line on the figure 11 shows the qualities for one generation. The best individual of each generation has an increasing quality value. Let's analyze the constructed features, for example: $\chi'_1 =$

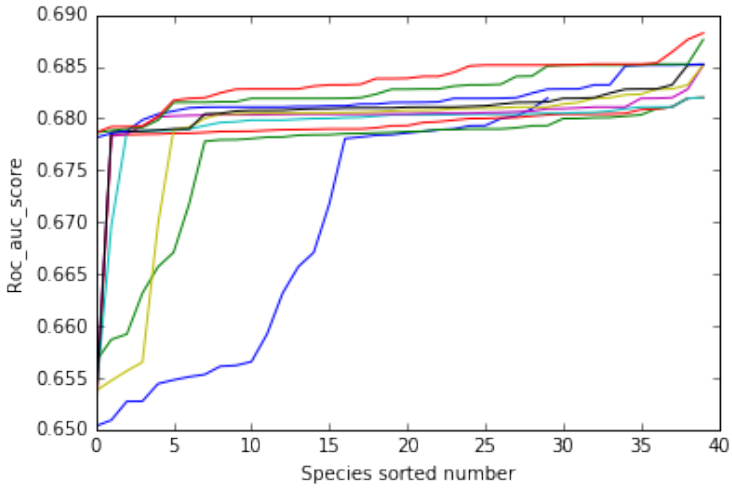

Рис. 11: The feature generation procedure convergence.

Model	Quality ($\mathcal{L}$) Roc-Auc	Quality ($\mathcal{T}$) Roc-Auc	Complexity
Logistic Regression	0.751	0.704	6
XGBoost	0.912	0.729	5000
NN	0.834	0.720	3000
Baseline	0.73	0.71	23
Proposed	0.782	0.730	25

Таблица 1: Quality comparing for German Credit Data

$\frac{1}{\chi_1+\chi_2+\chi_3}$, $\chi_2' = \sqrt{\chi_2 + \sqrt{\chi_1 + \chi_3}}$. Not all of these features are interpretable, there are inconsistencies in operations and dimensions: for example, the number of months at the current work is summed with the number of loans taken earlier is added up.

The solution is high interpretable. Table 1 compares the quality and complexity of models. The computantional experiment results show that the constructed solution is applicable, its quality remains at a high level with a low feature space complexity. Such a result was achieved due to the constructed procedure for generating and changing the feature space.

9 Conclusion

The paper proposes a unified procedure for converting a feature space for the credit scoring task. It combines known interpreted procedures for processing the feature space. The efficiency of the described approach is shown and the results of modeling on the open data are demonstrated. Continuation of work involves adding a procedure for feature selection

and complicating the feature generation process. Also, in future studies, it is planned to build a single adaptive online procedure for generating credit scoring multi-models over the resulting feature space.

СПИСОК ЛИТЕРАТУРЫ

1. Anne Kraus. Recent Methods from Statistics and Machine Learning for Credit Scoring // Dissertation an der Fakultat fur Mathematik, Informatik und Statistik der Ludwig–Maximilians–Universitat, Munchen. 2014.
2. Y. Dong. A Case Based Reasoning System for Customer Credit Scoring: Comparative Study of Similarity Measures // Proceedings of the 51st Annual Meeting of the ISSS. 2007. Tokyo, Japan.
3. S. V. Ulanov. Ocenka kachestva i sravnenie scoringovyh cart // Matematicheskie i instrumentalnye metody economiki. Economicheskie nauki. 9(58). 2009.
4. Ivan Oliveira, Manoj Chari, Susan Haller. Rigorous Constrained Optimized Binning for Credit Scoring // SAS Global Forum. 2008
5. Siddiqi N. Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring // New Jersey: John Wiley Sons. 2006.
6. USDOJ. Title VII—Equal Credit Opportunity Act // 2006. http://www.usdoj.gov/crt/housing/documents/ ecoafulltext_5- 1- 06.htm.
7. Sanja Scitovski and Natasa Sarlija. Cluster analysis in retail segmentation for credit scoring // 2014. CRORR 5. 235–245
8. L. J. Sanchez-Barrios, G. Andreeva, J. Ansell. Time-to-profit scorecards for revolving credit // European Journal of Operational Research. 2016. Vol. 249, Iss. 2. Pp. 397–406.
9. S. Maldonado, J. Pereza, C. Bravo. Cost-based feature selection for Support Vector Machines: An application in credit scoring // European Journal of Operational Research. 2017. Vol. 261, Iss. 2. Pp. 656–665.
10. C. Luo, D. Wu, D. Wu. A deep learning approach for credit scoring using credit default swaps // Engineering Applications of Artificial Intelligence. 2017. Vol. 65. Pp. 465–470.
11. X. Yufei, L. Chuanzhe, L. Y. Ying, L. Nana. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring // Expert Systems with Applications. 2017. Vol. 78. Pp. 225–241.
12. X. Chen, C. Zhou, X. Wang, Y. Li. The Credit Scoring Model Based on Logistic-BP-AdaBoost Algorithm and its Application in P2P Credit Platform //Proceedings of the Fourth International Forum on Decision Sciences. 2017. Pp. 119–130

13. G. Zeng. Invariant properties of logistic regression model in credit scoring under monotonic transformations // Communications in Statistics - Theory and Methods. 2017. Vol. 46. Iss. 17. Pp. 8791–8807.

14. Y. Wang, J. L. Priestley. Binary Classification on Past Due of Service Accounts using Logistic Regression and Decision Tree // Grey Literature from PhD Candidates. 2017. Vol. 4. http://digitalcommons.kennesaw.edu/dataphdgreylit/4

15. S. Walusala W, R. Rimiru, C. Otieno. A hybrid machine learning approach for cregit scoring using PCA and logistic regression // International Journal of Computer. 2017. Vol. 27. Iss. 1.

16. A. Mathew. Cregit scoring using logistic regression // San Jose State University. 2017. Master's Projects. 532. http://scholarworks.sjsu.edu/etd_projects/532

17. R. J. Connor. Grouping for Testing Trends in Categorical Data // Journal of the American Statistical Association. 1972. Vol. 67. Iss. 339.