Evolutionary Ensemble Approach for Behavioral Credit Scoring

Nikolay O. Nikitin(✉), Anna V. Kalyuzhnaya, Klavdiya Bochenina, Alexander A. Kudryashov, Amir Uteuov, Ivan Derevitskii, and Alexander V. Boukhanovsky

ITMO University, 49 Kronverksky Pr., St. Petersburg 197101, Russian Federation
nikolay.o.nikitin@gmail.com

Abstract. This paper is concerned with the question of potential quality of scoring models that can be achieved using not only application form data but also behavioral data extracted from the transactional datasets. The several model types and a different configuration of the ensembles were analyzed in a set of experiments. Another aim of the research is to prove the effectiveness of evolutionary optimization of an ensemble structure and use it to increase the quality of default prediction. The example of obtained results is presented using models for borrowers default prediction trained on the set of features (purchase amount, location, merchant category) extracted from a transactional dataset of bank customers.

Keywords: Credit scoring · Credit risk modeling · Financial behavior
Ensemble modeling · Evolutionary algorithms

1 Introduction

Scoring tasks and associated scoring models vary a lot depending on application area and objectives. For example, application form-based scoring [1] is used by lenders to decide which credit applicants are good or bad. Collection scoring techniques [2] are used for segmentation of defaulted borrowers to optimize debts recovery, and profit scoring approach [3] is used to estimate profit on specific credit product.

In this work, we consider scoring prediction problem for behavioral data in several aspects. First of all, the set of experiments were conducted to determine the potential quality of default prediction using different types of scoring models for the behavioral dataset. Then, the possible impact of the evolutionary approach to improving the quality of ensemble of different models by optimization of its structure was analyzed in comparison with the un-optimized ensemble.

This paper follows in Sect. 2 with a review of works in the same domain, in Sect. 3 we introduce the problem statement and the approaches for scoring task. Section 4 describes the dataset used as a case study and presents the conducted experiments. In Sect. 5 we provide the summary of results; conclusion and future ways of increasing the scoring model are placed in Sect. 6.

Y. Shi et al. (Eds.): ICCS 2018, LNCS 10862, pp. 825–831, 2018.
https://doi.org/10.1007/978-3-319-93713-7_81

2 Related Work

Credit scoring problem is usually considered within a framework of supervised learning. Thus, all common machine learning methods are used to deal with it: Bayesian methods, logistic regression, neural networks, k-nearest neighbor, etc. (a review of popular existing methods for credit scoring problems can be found in [4]).

A good and yet relatively simple solution to improve the predictive power of machine learning model is to use the ensemble methods. The key idea of this approach is to train different estimators (probably on different regions of input space) and then combine their predictions. In [5] authors performed a comparison of three common ensemble techniques on the datasets for credit scoring problem: bagging, boosting and stacking. Stacking and bagging on decision trees were reported as two best ensemble techniques. The Kaggle platform for machine learning competitions published a practical review of ensemble methods, illustrated on real word problems [6].

The current trend appears to be the enrichment of primary application form features with information about dynamics of financial and social behavior extracted from bank transactional bases and open data sources. According to some studies, the involvement of transactional data allows increasing the quality of scoring prediction significantly [7].

It is worth to mention that in the vast majority of the studies on credit scoring models are aimed to the resulting quality of classification and do not study in detail the possible effect of optimization the structure of ensemble of models. In contrast, in this paper we are aimed to investigate, how good the prediction for behavioral data can be and how much the ensemble structure and parameters can be evolutionary improved to increase the reliability of the scoring prediction.

3 Problem Statement and Approaches for Behavioral Scoring

The widely used approach for making a credit-granting (underwriting) decision is the application form-based scoring. It's based on demographic and static features like age, gender, education, employment history, financial statement, credit history. The application form data allows to create sufficiently effective scoring model, but this approach isn't possible in some cases. For example, pre-approved credit card proposal to debit card clients can be based only on a limited behavioral dataset, that bank can extract from the transactional history of a customer.

3.1 Predictive Models for Scoring Task

The credit default prediction result is binary, so the two-classes classification algorithms potentially applicable approach for this task. The set of behavioral characteristics of every client can be used as predictor variables, and default flag as the response variable. The several methods were chosen to build an ensemble: K-nearest neighbors classifier, linear (LDA) and quadratic separating surface (QDA) classifiers, feed-forward neural network with one single hidden layer, XGboost-based predictive model, Random Forest classifier.

An ensemble approach to scoring model allows to combine the advantages of different classifiers and to improve the quality of prediction of default. The probability vectors obtained at the output of each model included in the ensemble used as an input of a metamodel constructed using an algorithm that performs the maximization of the quality metrics and generates the optimal set of ensemble weights. While the total number of models combinations is 2^7, we implement the evolutionary algorithm using the DEoptim [8] package, which performs fast multidimensional optimization in weights space using the algorithm of differential evolution.

3.2 Metrics for Quality Estimation of Classification-Based Models

The standard accuracy metric is not suitable for the scoring task due to the very unbalanced sample of profiles with many "good" profiles (>97%) and a small amount of "bad." Therefore, threshold-independence metric AUC (the area under the receiver operating characteristic curve – ROC, that describes the diagnostic ability of a binary classifier) was selected to compare models.

Kolmogorov-Smirnov statistic allows estimating the value of threshold by comparison of probability distributions of original and predicted datasets. The probability which corresponds to the maximum of KS coefficient can be chosen as optimal in general case.

4 Case Study

4.1 Transactional Dataset

To provide the experiments with different configurations of scoring model, totally depersonalized transactional dataset was used. We obtained it for research purposes from one of the major banks in Russia. The dataset contains details for more than 10M anonymized transactions that were done by cardholders before they applied for a credit card and bank's application underwriting procedure accepted them. The time range of transactions starts on January 1, 2014, and covers the range up to December 31, 2016.

Each entity in the dataset is assigned to indicator variable of default, that corresponds to the payment delinquency for 90 or more days. The delinquency rate for the profiles from this dataset is 3.02%

The set of parameters of transactions included in the dataset and the summary of behavioral profile parameters that can be used as predictor variables in scoring models presented in Table 1.

This data allows identifying the profiles of bank clients as a set of some derived parameters, that characterize their financial behavior pattern, obtained from transactions structure. Also, the date variable can be used to take macroeconomic variability into account.

Since some profile variables have a lognormal distribution, the logarithmic transformation for one-side-restricted values and additional scaling to [0, 1] range was applied.

Table 1. Variables from the transactional dataset

Variable group	Transactional variables	Behavioral profile parameters
Attributes of transaction	IDs of client and contract, date of the transaction, date of contract signing	The numbers of actual and closed contracts
MCC (merchant category code)	Amount of transaction (in roubles), the location of The terminal used for operation (if known), transaction type (payment/cash withdrawal/ transfer), Merchant Category Code (if known)	Common frequency and quantitative characteristics of transactions; Merchant category-specific characteristics of transactions
Geo	Address of payment terminal	Spatial-based characteristics of transactions
Default mark	Binary flag of default	

4.2 Evolutionary Ensemble Model

The comparison of performance for scoring models is presented in Fig. 1.

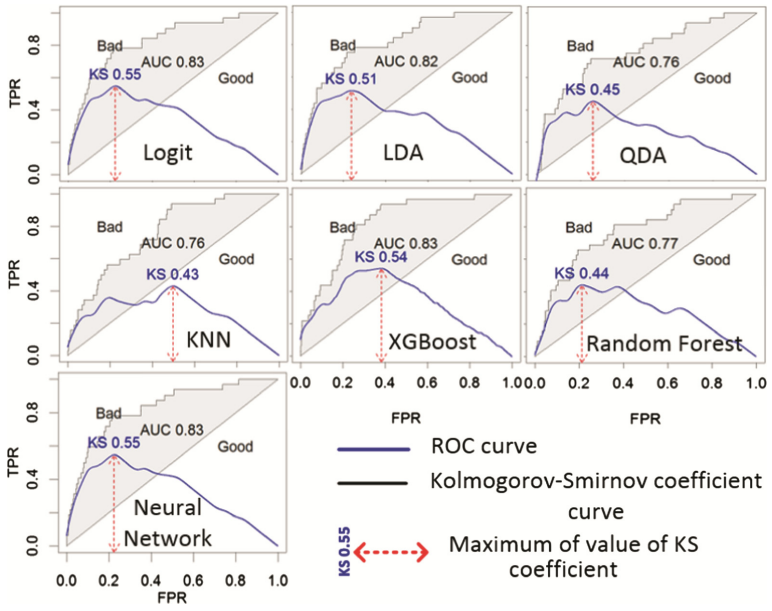

Fig. 1. The AUC and KS performance metrics for scoring models

The maximum value of Kolmogorov-Smirnov coefficient can be interpreted as an optimal value for probability threshold for each model.

The ensemble of these models can have a different configuration, and it's unnecessary to include all models to the ensemble. To measure the effect from every new model in the ensemble, we conduct a set of experiments and compare the quality of the scoring prediction for evolutionary-optimized ensembles with a different structure (from 2 to 7 models with separate optimization procedure for every size value). The logit regression was chosen as the base model. The structure of the ensemble is presented in Fig. 2a, the summary plot displaying the results is shown in Fig. 2b.

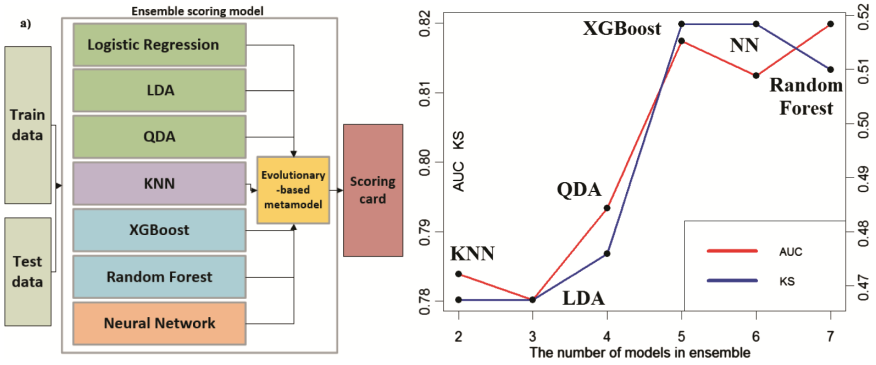

Fig. 2. (a) Structure of the ensemble of heterogeneous scoring models (b) Dependence of quality metrics AUC and KS on the number of models in optimized configuration the ensemble

It can be seen that the overall quality of the scoring score increases with the number of models used, but the useful effect isn't similar - for example, the neural network model does not enhance the ensemble quality.

The set of predictors of ensemble scoring models includes variables with different predictive power. The redundant variables make the development of interpretable scoring card difficult and can cause the re-training effect. Therefore, the evolutionary

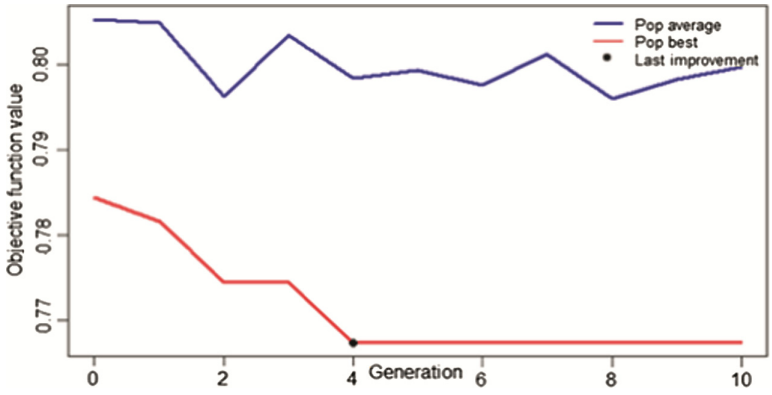

Fig. 3. The convergence of the evolutionary algorithm with an AUC-based fitness function

approach based on kofnGA algorithm [9] was used to create an optimal subset of input variables. The results of execution are presented in Fig. 3.

The convergence is achieved in 4 generations where experimentally determined population size equals 50 individuals; mutation probability equals 0.7 and variables subset size equals 14. The optimal variable set contains 8 variables from "MCC" group, 4 financial parameters and 2 geo parameters.

5 Results and Discussions

The experiments results can be interpreted as a confirmation of the effectiveness of evolutionary ensemble optimization approach. The summary of model results with 10-fold cross-validation and 70-30 train/test ratio is presented in Table 2.

Table 2. Summary of scoring models performance

Model	Training sample		Validation sample	
	KS	AUC	KS	AUC
Logistic regression	0.46	0.81	0.45	0.79
KNN	0.46	0.81	0.34	0.72
LDA	0.67	0.81	0.44	0.79
QDA	0.52	0.82	0.48	0.74
Random forest	0.97	0.99	0.41	0.78
XGBoost	0.6	0.88	0.49	0.81
Neural network	0.46	0.81	0.45	0.79
Base ensemble	0.74	0.94	0.49	0.80
Optimized ensemble	0.75	0.95	**0.51**	**0.82**

It can be seen that some best of single models provide similar quality of scoring predictions. The simple "blended" ensemble with equal weights for every model cannot improve the final quality, but the evolutionary optimization allows to increase the result of the scoring prediction slightly. The problem for the case study is the limited access to additional data (like applications forms), that's why the prognostic ability of the applied models can't be entirely disclosed.

6 Conclusion

The obtained results confirm that evolutionary-controlled ensembling of scoring models allows increasing the quality of default prediction. Nevertheless, it also can be seen that optimized ensemble slightly improve the result of the best single model (XGBoost) and, moreover, all results of individual models are relatively close to each other. This fact leads us to the conclusion that the further improvement of the developed model can be achieved by taking additional behavioral and non-behavioral factors into account to increase current quality threshold.

Acknowledgments. This research is financially supported by The Russian Science Foundation, Agreement № 17-71-30029 with co-financing of Bank Saint Petersburg.

References

1. Abdou, H.A., Pointon, J.: Credit scoring, statistical techniques and evaluation criteria: a review of the literature. Intell. Syst. Account. Fin. Manag. **18**(2–3), 59–88 (2011)
2. Ha, S.H.: Behavioral assessment of recoverable credit of retailer's customers. Inf. Sci. (Ny) **180**(19), 3703–3717 (2010)
3. Serrano-Cinca, C., Gutiérrez-Nieto, B.: The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. Decis. Support Syst. **89**, 113–122 (2016)
4. Lessmann, S., et al.: Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. Eur. J. Oper. Res. **247**(1), 124–136 (2015)
5. Wang, G., et al.: A comparative assessment of ensemble learning for credit scoring. Expert Syst. Appl. **38**(1), 223–230 (2011)
6. Kaggle Ensembling Guide [Electronic resource]
7. Westley, K., Theodore, I.: Transaction Scoring: Where Risk Meets Opportunity [Electronic resource]
8. Mullen, K.M., et al.: DEoptim: an R package for global optimization by differential evolution. J. Stat. Softw. **40**(6), 1–26 (2009)
9. Wolters, M.A.: A genetic algorithm for selection of fixed-size subsets with application to design problems. J. Stat. Softw. **68**(1), 1–18 (2015)