IEEE Second International Conference on Data Stream Mining & Processing
August 21-25, 2018, Lviv, Ukraine

Business-Oriented Feature Selection for Hybrid Classification Model of Credit Scoring

Galyna Chornous
Department of Economic Cybernetics
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
chornous@univ.kiev.ua

Ihor Nikolskyi
Department of Economic Cybernetics
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
ihor.nikolskyi@gmail.com

***Abstract* — Application of predictive models on the basis of data mining confirmed its expediency in solving many economic problems. One of the crucial issues is the assessment of the borrower's creditworthiness on the basis of credit scoring models. This paper proposed an ensemble-based technique combining selected base classification models with business-specific feature selection add-on to increase the classification accuracy of real-life case of credit scoring. As the model limitations have been used easy-understandable algorithms on open-source software (R programming). The statistical results proved that hybrid approach for user-defined variables can be more than useful for ensemble binary classification model. It is shown that a great improvement can be reached by applying hybrid approach to feature selection process on additional variables (more descriptive ones that were built on initial features) for this real-life case with limited computational resources.**

Keywords — hybrid, ensemble, feature selection, binary classification, stacking, major voting, credit scoring, R programming

I. INTRODUCTION

Application of predictive models on the basis of data mining (DM) confirmed its expediency in solving many economic problems: forecasting changes in stock indexes, price and production management, analysis of insurance risks, diagnostics of bankruptcy, credit card fraud, etc.

One of the crucial issues is the assessment of the borrower's creditworthiness on the basis of credit scoring models. The approach advantage is to avoid subjectivity in making credit decisions, to justify relevant decisions based on the knowledge extracted through the use of intelligent analysis of the accumulated data sets of borrowers.

A lot of statistical and machine learning techniques such as Logistic Regression, Support Vector Machines, Bayesian Networks, Neural Networks and Decision Tree algorithms are common tools for measuring creditworthiness [1; 2; 3].

The latest researches demonstrate that the implementation of separate stand-alone models for solving complex tasks does not always lead to success. Currently, the search for "strong" individual model is no longer relevant for most of researchers: the point of interest is about large ensembles of "weak" methods and algorithms [4; 5; 6].

The integration of various methods gives opportunity to deal with the limitations of each individual method, and in turn provides new opportunities for supporting the decision-making process within a single architecture and leads to the creation of hybrid intelligence systems (HIS). The methodological support of HIS is based on a combination of different methods of AI, operations research, decision making theory and system analysis, the formation of ensemble models and the use of hybrid algorithms [7].

Forecasting the borrower's creditworthiness is one of the typical tasks of DM, it involves a binary classification of borrowers. To support the solution of the classification problem, the researchers offer a modern, complex implementation of methods and models such as advanced machine learning tools, such as ensembles of classifiers and hybrid classifiers, provide higher accuracy and performance. However, not all companies have proper human and computational resource for developing and implementation of state-of-the-art solutions. It limits the power of data science, i.e. a live issue is to find a way to improve performance of free common open-source solutions.

That’s why the idea of this work is to find a way to dramatically improve model performance with limited computational resources, which is typical for Ukrainian business. The idea is to build a hybrid feature selection add-on for a typical ensemble model for binary classification.

II. RELATED WORKS

The assessment of the borrower's creditworthiness is one of the crucial issues for the financial institutions. The credit classification scoring is a great technique for understanding the risk of individual borrowers, gauging overall risk exposure and building data-driven, risk-adjusted strategies for existing customers. Therefore, great attention of researchers is paid to this issue in recent years.

We have studied a lot of publications regarding performance evaluation of DM algorithms on different tools, that were presented during the last 3 years. Some of them are described below.

L. Zernova [8] studied the relationship between the concepts of individual borrower risk, creditworthiness, the techniques of estimating creditworthiness, refined the concept and factors of creditworthiness, developed a methodology of scoring. N. Siddiqi [9] showed the most recent trends that widen credit scoring functionality and new in-depth analyses. It deals with defining infrastructure for in-house credit scoring, validation, governance, and Big Data. The authors [10] overview ideas of the statistical and operations research methods used in building credit and behavioral scorecards, as well as the advantages and disadvantages of each approach. These studies do not include specific developments in the hybrid approach, but present some examples of them.

H. Chen etc. [11] advanced Bayesian algorithm for credit assessment. The new trial ensembles logistic regression analysis (LRA), cluster and MLP-NN in Bayesian approach as an advanced classifier. S. Dahiya etc. [12] proposed an hybrid modeling technique using seven individual models (the NNs, C5.1, CART Tree, QUEST, CHAID, LR and SVM) to increase the model performance. Feature selection has also been used for selecting important attributes for classification. In this particular case Chi-Square statistic was adopted for choosing the most important ones out of all basic features. A.G. Armaki etc. [13] combined traditional and ensemble methods and come up with a hybrid meta-learner model The structure of the model is based on the traditional hybrid model of 'classification + clustering'. They propose several versions of the hybrid model by using various combinations of classification and clustering algorithms. S.H. Van etc. [14] built a creditworthiness classification model based on parallel Gradient Boosted Model, filter and wrapper approaches to estimate the credit score from the input features. Selected scoring variables are combined by feature importance (Gini index) and Information Value. H. Xiao etc. [15] proposed a hybrid classification method based on supervised clustering for credit scoring. Clusters from different classes are then pairwise combined to form a number of training subsets. In each training subset, a specific base classifier is constructed. M. Ala'raj and M.F. Abbod [16] presents a new hybrid ensemble credit scoring model through the combination of two data pre-processing methods based on Gabriel Neighbourhood Graph editing and Multivariate Adaptive Regression Splines in the hybrid modelling phase.

Despite the generous amount of studies and high quality results, the issue of which classifier is the best remains open, since it is very difficult to develop an all-in-one model that adapts to each set of data or set of attributes. Especially in the face of a shortage of human and computational resources for the implementation of the most up-to-date solutions.

III. METHODOLOGY

A. Dataset and tools

The main task of this work is the construction of an algorithm for binary classification in the context of determining the creditworthiness by personal data.

The data was collected by Dream Housing Finance Company in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan. The given problem is to identify the customer segments, those are eligible for loan amount so that they can specifically target these customers. The data was collected for Analytics Vidhya loan prediction hackathon [17]. Dataset consists of 614 observations and the structure is presented in the Table I. The data was collected from online registration forms.

To perform all stages of the experiment one of the most popular DM tools was chosen – R programming. R is an efficient tool for evaluating great variety of DM techniques such as preprocessing, classification, clustering etc. It for was used for all stages of the experiment: data retrieving, data cleaning, feature selection, model building and validating the results. The base libraries are caret and caretEnsemble.

TABLE I. DATASET STRUCTURE

Variable	Type	Description
Loan_ID	Factor	Applicant ID
Gender	Factor	Gender
Married	Factor	Marital status
Dependents	Numeric	Number of dependents
Education	Factor	Education status
Self_Employed	Factor	Employment status
ApplicantIncome	Numeric	Income level
CoapplicantIncome	Numeric	Co-applicant income level
LoanAmount	Numeric	Asked loan
Loan_Amount_Term	Numeric	Payout period
Credit_History	Factor	Positive or negative history
Property_Area	Factor	Residence status
Loan_Status	Factor	Approved / disapproved

B. Concepts used

The concepts used in the experiment are feature selection, classification and ensembled modeling (ensemble stacking).

1) Feature selection

Feature selection or selection of information attributes is the process of selecting the most significant features for their further use in machine learning and statistics. Selection of attributes helps to solve the following tasks:

- Avoid overtraining the model by removing irrelevant characteristics.
- Avoid the "Curse of Dimension".
- Simplify the model and reduce the time for its training by removing redundant or insignificant features from the input data.

It is important to distinguish the feature selection from the dimensionality reduction. In the second case, the set of originally collected features is transformed into another set, which is more suitable for machine learning.

In this experiment Information Gain, Chi-Squared and Mean Decrease Gini methods.

- *Information Gain*. The entropy characterizes the impurity of an arbitrary collection of samples. Information Gain is the expected reduction in entropy caused by partitioning the samples according to a given feature which is the way of measuring association between inputs and outputs.
- *Chi-Squared test*. Pearson's chi-squared test is a common statistical tool used for categorical data to understand the probability that any observed difference between the sets arose by chance and to test the independence of two events. it is used to find whether the occurrence of a certain feature and the occurrence of a specific class are independent. Thus, after estimating the following quantity for each feature they are ranked by the score. Higher scores states that the null hypothesis (H0) of autonomy is not confirmed which means that the occurrence of the term and class are dependent.
- *Mean Decrease Gini coefficient.* Mean decrease in the Gini impurity criterion is used as a way to estimate variable importance. The concept implies that every time a feature is used to split a node, the Gini coefficient for the child nodes are calculated to be

compared with the original node. The greater decreases, the stronger relationship with the output [18].

2) Feature engineereenig

The concept of feature engineering is to construct the process of using given knowledge of the data to select or build new features which are more suitable for machine learning algorithms. Feature engineering is a necessary stage of machine learning. It requires both analytical creativity and powerful computational resources. However, a good feature selection add-on can be automated through a combination of basic routine methods.

Feature engineering is more about human side of machine learning, but its usage is essential in applied machine learning. The basic process of feature engineering consists of the following steps:

- Consolidating fundamental ideas about feature set;
- Creating concepts for new features;
- Implementation of the ideas to a given dataset;
- Testing model sensitivity and performance with new features;
- Adjusting new features;
- Looping through this process until the best combination of features is found.

3) Binary Classification

Today binary classification is a widely used method in such spheres as medicine, telecommunication, economics, trading, sociology etc.

Binary classification is a common task in machine learning universe. This is an example of supervised learning technique, a method of machine learning where the classes are known beforehand and is used to classify new observations into predefined categories. Speaking about binary classification, there are only two classes.

The task of binary (binomial) classification is classifying the given elements into two groups (predicting which group each one belongs to) using a classification engine. A set of parameters (numeric, factor etc.) define the context of each observation which is an input for classification rule.

Accuracy metrics in term of binary classification engines calculate the two classes of correct predictions and two classes of errors. Typical metrics are accuracy (ACC), precision, recall, false positive rate, F1-measure. Every metric evaluates a certain aspect of the forecasting model. Accuracy (ACC) estimates the share of correct predictions. Precision estimates the share of true positives predictions among all observations that are predicted as positive. Recall estimates how many true positives were predicted as positive. F-measure is the harmonic mean of precision and recall.

Typical methods for performing binary classification are Decision Trees, Random Forest, Support Vector Machines, Neural Networks, Logistic Regression [5].

4) Stacking ensemble

Simultaneous global optimization of all basic algorithms used to construct the ensembles is a complex multicriteria problem. Such an optimization requires knowledge of the internal structure of the algorithms, which complicates the use of standard teaching methods. In practice, the implementation of such a strategy is embodied in certain improvements of consistent and parallel approaches.

One of the corresponding implementations is level aggregation or stacking. Unlike bagging and booting, stacking is usually not used to combine models of the same type but applies to models built using various learning algorithms. Staking tries to learn each model using a meta-learning algorithm, which allows to find the best combination of outputs of the basic algorithms [19] (Fig. 1).

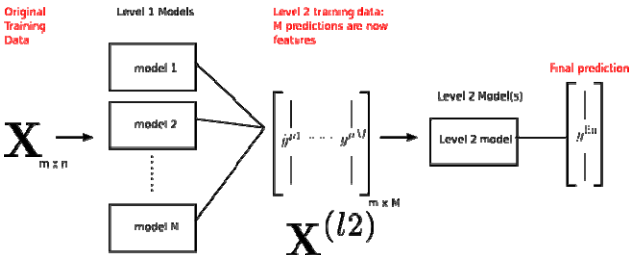

Fig. 1. Basic concept of stacking ensemble

Work with the ensemble includes the choice of basic neural network architecture and the number of networks in the ensemble, training (with the help of, for example, bagging or other algorithm), work with the received model (data processing, copying / serialization of the model, etc.).

The ensembles are widely used in practice, especially in predicting bankruptcy, determining credit scores etc. Examples of the implementation of ensembles are given in the studies.

IV. Experimental Results

A. Proposed experiment

As stated above the key of the experiment is to improve performance of a multicriterial ensembled model for solving the task of binary classification on the case of determining the creditworthiness by personal data. Model concept is presented in Fig. 2.

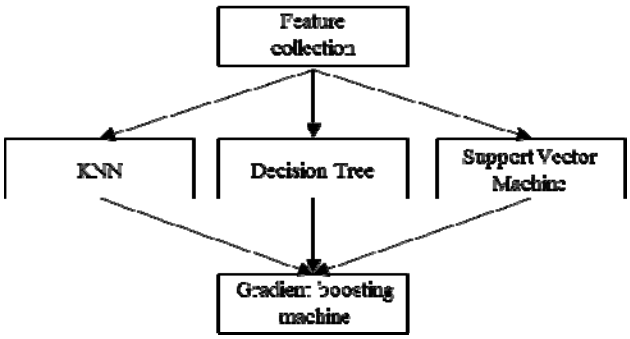

Fig. 2. Proposed model

The different steps that were carried out for conducting the experiment have been shown in this section. The high-level description of framework used to build the hybrid model is given in the Fig. 3.

To make the paper be outstanding form the variety of models for binary classification the following steps have been made:

- Applying machine learning to real-life case with limitations of Ukrainian business reality.
- Including heuristic analysis for data cleaning process to consider the origin of data
- Adding new features, which are more business informative than initial ones.
- Combination of several different feature selection methods, which gives the ability to correctly process both factor and numeric variables.
- Improving performance of common algorithms by combination of technical and analytical approaches.

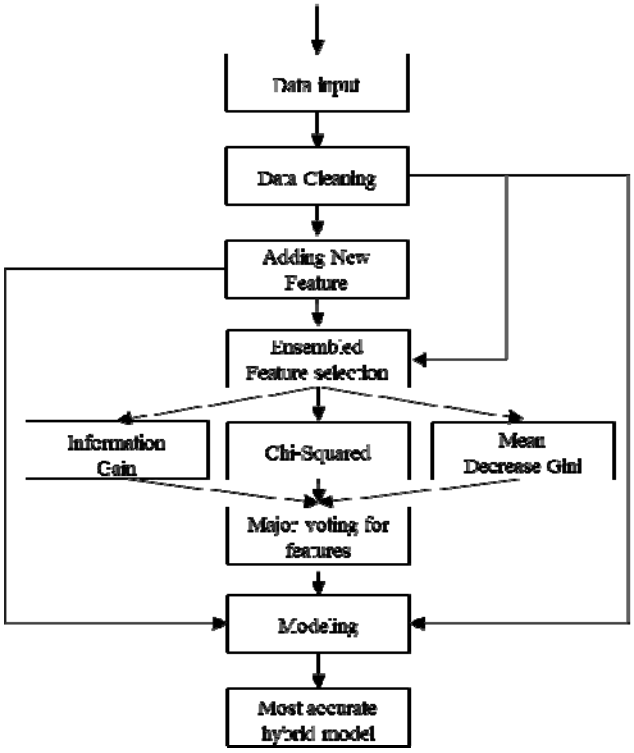

Fig. 3. Proposed architecture for the experiment.

According to the experiment plan the first tasks were about retrieving and cleaning data. As the idea of experiment is also aimed at thorough data cleaning, the various data preprocessing techniques were performed to remove noisy and redundant data from the database. The procedure of cleaning included removing observations with empty features and extreme values, filling missed values with median (for numeric features) and most popular (factor features) values. Exceptional approach was applied to features "Married" and "Self_Employed" ("No" if missed) due the nature of these fields.

Feature selection is also a data preprocessing technique to select the relevant attributes for the experiment. Feature engineering is crucial for model optimization.

The modeling stage implies looping through different combinations of features used for binary classification to find the most accurate approach.

B. Statistics results

This section presents the statistical results of all stages of the experiment. The underlying idea is to adjust business logic to the process of feature selection by adding new feature interaction rules, which are presented in the Table II.

The next step was to apply hybrid approach for feature selection by major voting of mean decrease Gini, information gain and Chi-squared coefficient methods. The statistical results are presented in the Table III. In case if modified features outperform basic ones, basic variables are excluded from the collection of important features.

TABLE II. FEATURE INTERACTION RULES

New features	Interaction
IncomePerPerson	=ApplicantIncome / (Dependents+1)
MonthlyPayment	=Loan_Amount / Loan_Amount_Term
Friend	=if(CoapplicantIncome>0; True; False)
Family_backup	=if(AND(CoapplicantIncome>0;Married="Yes"); True;False)

The results in the Table IV show that modified features almost all tend to describe the nature of data better and be more informative in terms of data. So, the target collection of important features consists of Credit_History, Friend, Family_backup, MonthlyPayment, IncomePerPerson.

The next table describes comparison of basic methods for classification with three feature collections: basic features, all features and important features.

As we see, the main hypothesis is proven: ensembled model can benefit from hybrid approach for feature selection, especially with analytically defined variables, which results in slightly higher accuracy of a model.

TABLE III. FEATURE INTERACTION RULES

Feature	Importance	Information gain	Chi.Squared coefficient
Credit_History	**0,039**	**0,099**	**0,457**
Friend	**0,037**	**0,099**	**0,457**
Family_backup	**0,022**	**0,062**	**0,346**
ApplicantIncome	0,019	0,006	0,113
MonthlyPayment	**0,018**	**0,005**	**0,105**
LoanAmount	0,018	0,003	0,076
IncomePerPerson	**0,015**	**0,001**	**0,037**
CoapplicantIncome	0,012	0	0,029
Married, Dependents, Gender, Self_Employed, Loan_Amount_Term, Education, Property_Area	< 0,01	0	0

TABLE IV. EVALUATION OF MODEL PERFORMANCE

	Accuracy	F-measure
All initial features	0,692	0,679
All features	0,674	0,671
Ensembled approach		
Important initial features	0,776	0,753
Important modified features	**0,841**	**0,832**

V. CONCLUSION

This paper proposes an ensemble-based technique combining selected base classification models with business-

specific feature selection add-on to increase the classification accuracy of real-life case of credit scoring.

The model limitations were to use commonplace easy-understandable algorithms on open-source software (R programming).

The key points of this study are applying machine learning to real-life case with limitations of Ukrainian business reality; including heuristic analysis for data cleaning process to consider the origin of data; adding new features, which are more business informative than initial ones; combination of several different feature selection methods, which gives the ability to correctly process both factor and numeric variables; improving performance of common algorithms by combination of technical and analytical approaches.

The statistical results proved that hybrid approach for user-defined variables can be more than useful for ensembled binary classification model. That means that a great improvement can be reached by applying hybrid approach to feature selection process on additional variables (more descriptive ones that were built on initial features) for this real-life case with limited computational resources.

REFERENCES

[1] A.Q. Kadhim, G.A. El-Refae, and S.F. El-Itter, "Neural Networks in Bank Insolvency Prediction," International Journal of Computer Science and Network Security, vol 10, no 5, pp. 240–245, 2010.

[2] T. Pavlenko, and O. Chernyak, "Credit risk modeling using bayesian networks," International Journal of Intelligent Systems, vol. 25, issue 4, pp.326–344, 2010.

[3] G.O. Chornous, Proactive Management of Socio-Economic Systems Based on Intellectual Data Analysis: Methodology and Models. Kyiv: Kyiv University, 2014.

[4] M. Kim, and D. Kang, "Ensemble with neural networks for bankruptcy prediction," Expert Systems with Applications, vol. 37, issue 4, pp. 3373–3379, 2010.

[5] C.-F. Tsau, and J.-W. Wu, "Using neural network ensembles for bankruptcy prediction and credit scoring," www.sciencedirect.com

[6] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms. New York: Wiley and Sons, 2014.

[7] L.R. Medsker, Hybrid Intelligent Systems. Boston: Springer, 2013.

[8] L. Zernova, The creditworthiness of bank's clients: Analysis and assessment. LAP LAMBERT Academic Publishing, 2016

[9] N. Siddiqi, Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards, 2nd ed.. Wiley, 2017.

[10] L. Thomas, J. Crook, and D. Edelman, Credit Scoring and Its Applications, 2nd Revised ed. SIAM-Society for Industrial & Applied Mathematics, 2017.

[11] H. Chen, M. Jiang, and X. Wang, "Bayesian Ensemble Assessment for Credit Scoring," 4th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS), 2017.

[12] S. Dahiya, S.S. Handa, and N.P. Singh, "Credit Scoring Using Ensemble of Various Classifiers on Reduced Feature Set," Industrija, vol.43, no.4, pp. 163-172, 2015.

[13] A. G. Armaki, M. F. Fallah, M. Alborzi, and A. Mohammadzadeh, "A Hybrid Meta-Learner Technique for Credit Scoring of Banks' Customers," Engineering, Technology & Applied Science Research, vol. 7, no. 5, pp. 2073-2082, 2017.

[14] S. H. Van, N. N. Ha, and H. N. Thi Bao, "A hybrid feature selection method for credit scoring," EAI Endorsed Transactions on Context-aware Systems and Applications, vol. 4, issue 1, 09 2016 - 03 2017.

[15] H. Xiao, Z. Xiao, and Y. Wang, "Ensemble classification based on supervised clustering for credit scoring," Applied Soft Computing, vol. 43, pp. 73-86, June 2016.

[16] M. Ala'raj and M. F. Abbod, "A new hybrid ensemble credit scoring model based on classifiers consensus system approach," Expert Systems with Applications, vol. 64, pp. 36-55, December 2016.

[17] Analytics Vidhya / Loan Prediction: Practice Problem // https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/

[18] K. S. Cho, "Ensemble learning with feature selection for Alzheimer's disease prediction," – http://www.academia.edu/30496678, 2016.

[19] B. Himmetoglu, "Stacking models for improved predictions" – https://burakhimmetoglu.com/2016/12/01/stacking-models-for-improved-predictions/, 2017