A hybrid approach for feature selection in data mining modeling of credit scoring

Galyna O. Chornous[1], Kostiantyn K. Pysanets[1], Nataliia O. Yakovenko[1]

[1]Taras Shevchenko National University of Kyiv, 90a, Vasylkivska str., Kyiv, 03022 Ukraine
chornous@univ.kiev.ua, knukkp@gmail.com,
tasha.yakovenko@gmail.com

Abstract. Recent year researches shows that data mining techniques can be implemented in broad areas of the economy and, in particular, in the banking sector. One of the most burning issues banks face is the problem of non-repayment of loans by the population that related to credit scoring problem. The main goal of this paper is to show the importance of applying feature selection in data mining modeling of credit scoring. The study shows processes of data pre-processing, feature creation and feature selection that can be applicable in real-life business situations for binary classification problems by using nodes from IBM SPSS Modeler. Results have proved that application of hybrid model of feature selection, which allows to obtain the optimal number of features, conduces in credit scoring accuracy increase. Proposed hybrid model comparing to expert judgmental approach performs in harder explanation but shows better accuracy and flexibility of factors selection which is advantage in fast changing market.

Keywords: Credit Scoring Model, Feature Selection, Hybrid Approach, Data Mining, IBM SPSS Modeler

1 Introduction

Recent year researches shows that data mining techniques can be implemented in broad areas of the economy and, in particular, in the banking sector. Banks and other credit institutions have faced the need to process large amounts of data at a growing rate. The imperatives for the volume of data operations and the speed of their processing require these processes to be almost completely automated. These requirements apply not only to direct digitalization, but also to the procedures for developing appropriate mathematical models. Credit scoring models are a prime example. They are increasingly combined with new computational methods based on data mining.

An extremely important problem in scoring modeling was and remains the choice of the borrowers' characteristics, which are decisive in loan decision making. In terms of the model, these characteristics are often known as the explanatory variables, covariates, predictor attributes, predictor variables, independent variables or, typically, features. Set of the most influential features is not permanent. It changes over time and is significantly dependent on the macroeconomic situation and national specificities.

Researchers [16] analyzed 187 papers from 1992 to 2015 on credit scoring and noted that feature selection is the No.4 objective of seven types of main objectives: proposing a new method for rating, comparing traditional techniques, conceptual discussions, feature selection, literature review, performance measures studies and other issues. The authors of this research have set this objective for 95 articles, representing 51% of the total, with 52 articles published since 2011. Among the main studies of the 2002-2015 period, which are devoted to the feature selection, can be highlighted [2, 4, 6, 11, 14, 15, 18, 20, 22, 27, 28].

Relevant problems are also actively researched in the last four years. The examples of publications are as follows [2, 3, 8, 10, 17, 24-26]. The focus of pertinent researches on feature selection is increasingly being shifted toward machine learning and hybridization methods.

Nowadays software implementations of machine learning algorithms in credit scoring can take place using the following classes of software: business application packages (statistical packages and analytics platforms, such as SAS/STAT, SAS Enterprise Miner, IBM SPSS Modeler, STATISTICA Data Miner), open platforms (Python, R, Apache Spark) and cloud solutions (Microsoft Azure Machine Learning Studio, Google Machine Learning Engine).

Despite all the advantages of cloud-based data analytics, it is not widely used in banks due to security concerns about the passing confidential data to the cloud.

The biggest advantage of open environments is the ability to use much more algorithms comparing to business application packages. However, they impose additional qualification requirements for the developers of scoring models.

Therefore, statistical packages and analytical platforms are the most commonly used in banks. They include analytical pre-proceeding tools, ready and customizable machine learning algorithm templates. In addition, these packages allow to configure model settings and use interactive quality assessment techniques. This conclusion is also confirmed by the authors of the paper [26].

Business application packages have powerful functionality to solve the problem of feature selection, which can be improved by combining built-in approaches and hybridization.

That is why, it is extremely important to improve the functionality of these software products for developing scoring models in general, and for making feature selection in particular.

The purpose of this study is to propose new hybrid approaches to the feature selection, which will improve the quality of credit scoring models, built on intelligent data analysis, machine-learning approaches in today's analytics platforms.

2 Literature review

From last few decades more and more attention has been paid to the problem of credit scoring [21, 23]. Artificial Neural Networks (ANNs) [25] and Support Vector Machine (SVM) [4, 20] are two commonly soft computing methods used in credit scoring modelling. Other methods like evolutionary algorithms, stochastic optimization technique have shown promising results in terms of prediction accuracy [26]. Besides, there are also traditional approaches based on expert knowledge which allow to develop expert judgmental models [7], scoring expert systems [1] and mixed models.

Feature selection algorithms, generally as preprocessing methods of scoring model creation, can be used to increase the classification performance. They have a number of benefits as follows [19]: decreasing the noise in dataset; reducing the computational cost in order to successfully acquire proper models; helping to better understand the final models in the classification algorithms; simple application; assisting in updating the model.

We have studied a lot of publications regarding the problem of credit scoring in data mining with applying feature selection techniques. Some of them are described below.

Starting with primitive approaches, use of expert judgmental forms is a good source for initial features list creation. It is not common for long existing credit business, however in case of new lending segments emergence under condition of data shortage it shows acceptable effectiveness.

One of the simplest and widespread statistical approach is ‘weight of evidence’ and ‘information value’ indicators use, explained by Siddiqi in [21].

Kuhn and Johnson describe in [12-13] two main types of feature selection techniques: wrapper and filter methods. The filter approach considers the feature selection process as a separate step of learning algorithms. The filter model uses evaluation functions to evaluate the classification performances of subsets of features. There are many evaluation functions such as feature importance, Gini, information gain, the ratio of information gain, etc. A disadvantage of this approach is that there is no relationship between the feature selection process and the performance of learning algorithms.

The wrapper approach uses a machine-learning algorithm to measure the set goodness of selected features. The measurement relies on the performance of the learning algorithm such as its accuracy, recall and precision values.

The papers [25-26] systemize a credit scoring model based on deep learning and feature selection to evaluate the applicant’s credit score from the applicant’s input features.

The objective of many studies is to analyze the outperform feature selection techniques among conventional and heuristic techniques in various applications [14, 17, 28].

A lot of researches embody the optimization approach to find the best subset of predictors for improving scoring model performance [3, 6, 27]. For instance, in [3] authors suggested to study local search, stochastic local search and variable neighborhood search for feature selection in credit scoring. The proposed feature selection is then combined with a support vector machine to classify the input data.

Publications of the recent years show for credit scoring problem active use of principal component analysis feature selection: PCA is a transformation process to reduce the number of features by extraction of the new independent features [7, 11, 25].

These days, there are more and more examples of the use of different hybrid feature selection techniques.

A hybrid approach in data mining models of feature selection algorithms and ensembling learning classifiers for credit scoring was used by Koutanaei, Sajedi and Khanbabaei [11]. Credit scoring modeling based on feature selection approach and parallel Random Forest were described by Ha Van Sang, Ha Nam and Nguyen Duc Nhan [8]. For feature selection, a feature clustering approach was proposed to find optimal set of predictors by autors [24]. Kamalov and Thabtah suggested a new filtering method that combines and normalizes the scores of three major feature selection methods: information gain, chi-squared statistic and inter-correlation [10]. Chornous and Nikolskyi prove 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 case with limited computational resources [5].

This study in comparison with others proposes a unique approach for feature selection techniques and has in advantage the accessibility of the approach for users of analytical platforms, business application packages, using the capabilities of the built-in tools and offering methods for combining them.

Scoring modelling techniques mentioned in studied works are commonly used, but their aim mostly to show an increase of models performance accuracy directly. Previous studies compare different approaches or miss the relationship between the feature selection process and the performance of learning algorithms. So the authors' aim is the search of techniques conjunction for features selection to reach and explain better models performance.

Besides, it is obvious that the usage of feature selection techniques differs among countries, so the results of this particular study can be implemented in real business cases in Ukraine.

Important tasks of this research are to show the advantages of modern powerful analytical platforms (on the example of IBM SPSS Modeler) for solving the problems of credit scoring in general and making the feature selection in particular; to suggest the concept of an effective combination of their tools; to show the experimental results of the joint application of options in Feature Select node and PCA/Factor node to optimize the feature selection process and to model credit scoring on the example of Ukrainian bank data.

3 Methodology

Dataset and tools

The dataset was collected and systematized all socio-economic information about them at the stages of loan repayment, collected information about the timeliness of the loan by Ukrainian bank during the stages of providing consumer loans to individuals.

The original dataset consists of 61 fields with a record volume of 61 216. It is advisable to use all the data in this database, not just a subset, which will allow us to build more accurate models. Non-relevant records or non-relevant attributes may not be included. The available data is in several formats: numeric, categorical and logical.

All the computing work was done in nodes from IBM SPSS Modeler. This software product has powerful functionality for solving binary classification tasks, including credit scoring.

To select important features, such built-in tools as the Feature Selection node and the PCA/Factor node are used [9].

The Feature Selection node allows to implement 3 key procedures:

1. Screening that removes unimportant and problematic inputs and records, or cases such as input fields with too many missing values or with too much or too little variation to be useful.
2. Ranking that sorts remaining inputs and assigns ranks based on importance.
3. Selecting that identifies the subset of features to use in subsequent models.

The PCA/Factor node provides powerful data-reduction techniques to reduce complexity of data. Two approaches are provided. The first one is Principal component analysis (PCA). In this statistical dimensionality reduction technique, the correlated features can be combined as principal components. The second one is Factor analysis. It identifies underlying concepts, or factors, that explain the pattern of correlations within a set of observed fields. Factor analysis focuses on shared variance only. Variance that is unique to specific features is not considered in estimating the model. Several methods of factor analysis are provided by the Factor/PCA node. For extended comparison we have received expert judgmental generic list of influencing factors, their attributes and weights for the studied segment from one of professionals of Ukraine credit market with more than 10 years of working experience.

For both approaches, the goal is to find a small number of derived features that effectively summarize the information in the original set of features and compare them with expert model.

To develop scoring models, IBM SPSS Modeler offers 16 base methods (some examples of these ones are: Decision Trees (CART, QUEST, C.5.0, CHAID), Neural Network, SVM, Bayes Network, KNN, Logistic Regression, Discriminant analysis) and large set of ensemble methods (bagging, boosting, Random Tree, Random Forest, XGBoost Tree XGBoost Linear, XGBoost-AS) [9].

Moreover, the Auto Classifier node creates and compares a number of different binary models, allowing to choose the best approach for development. 16 modeling algorithms are supported, making it possible to select the best methods, the specific options for each, and the criteria for comparing the results.

Concept used

Our investigation starts with the stages of pre-processing the dataset and adding new features that better describes defining borrower’s status than initial ones. Than to the resulted dataset with initial number of features we applied modeling methods such as Decision Trees, Random Forest, Support Vector Machines, Neural Networks, Logistic Regression and Expert approach. After this stage we applied a number of feature selection techniques in order to decrease a number of features and conduct modeling again but under feature selection. We provide with analysis to compare results for AUC values between initial models and models under feature selection. Finally, we applied a hybrid approach of feature selection analysis to obtain the optimal number of features, conduces in credit scoring accuracy increase. Described stages for investigational results are presented in Figure 1.

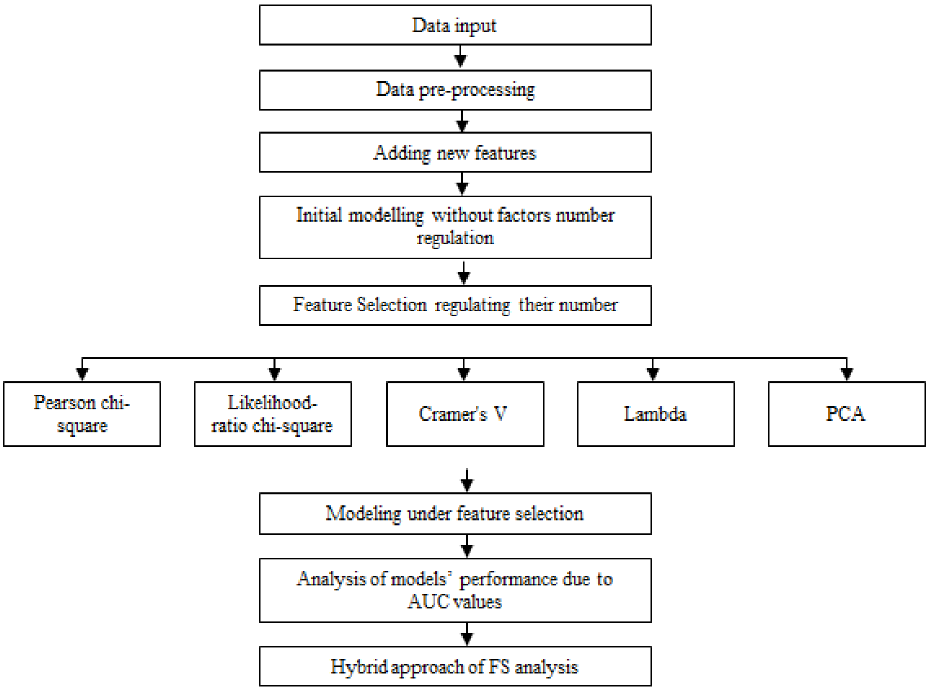
Data input
Data pre-processing
Adding new features
Initial modelling without factors number regulation
Feature Selection regulating their number
Pearson chi-square
Likelihood-ratio chi-square
Cramer's V
Lambda
PCA
Modeling under feature selection
Analysis of models' performance due to AUC values
Hybrid approach of FS analysis

Fig. 1. Proposed concept for the experiment

Data pre-processing

Data pre-processing includes several steps:

1. Removing missing values and irrelevant features.
2. Data integration, transformation and normalization.
3. Reclassifying categorical values.
4. Balancing data.

Feature creation

The process of feature creation gives an opportunity to adjust business logic to the process of feature selection by adding new feature interaction rules. Adding new features gives more description for defining borrower's status than initial ones, should increase modeling results in order to improve banks' performance in credit scoring problem.

Feature selection

Feature selection is also a data pre-processing technique, which is used to select the relevant attributes for the experiment. Feature engineering is crucial for model optimization.

This paper proposes feature selection by ranking measures as Pearson chi-square, Likelihood-ratio chi-square, Cramer's V and Lambda and feature selection using Principal Component Analysis. All feature selection techniques in detail are presented below.

Feature selection by ranking measures.
Feature selection by ranking measures was used to screen and rank features by importance. In this paper, we focus on 4 ranking measures.

1. Pearson chi-square. Tests for independence of the target and the input without indicating the strength or direction of any existing relationship.
2. Likelihood-ratio chi-square. Similar to Pearson's chi-square but also tests for target-input independence.
3. Cramer's V. A measure of association based on Pearson's chi-square statistic. Values range from 0, which indicates no association, to 1, which indicates perfect association.
4. Lambda. A measure of association reflecting the proportional reduction in error when the variable is used to predict the target value. A value of 1 indicates that the input field perfectly predicts the target, while a value of 0 means the input provides no useful information about the target.

Feature selection by Principal Component Analysis.
Principal components analysis (PCA) finds linear combinations of the input fields that do the best job of capturing the variance in the entire set of features, where the components are orthogonal (perpendicular) to each other. PCA focuses on all variance, including both shared and unique variances.

Factor analysis and PCA can effectively reduce the complexity of data without sacrificing much of the information content. These techniques can help to build more robust models that execute faster than would be possible with the raw input fields.

Modeling

Typical methods for performing binary classification are Decision Trees, Random Forest, Support Vector Machines, Neural Networks, Logistic Regression [23]. Expert approach is an effective alternative for these methods.

In this paper, we focus on the four main methods as Support Vector Machines, Neural Networks, Logistic Regression and Decision Tree (CHAID). We are interested in achieving the best rate of AUC, because the higher is the value of AUC the better is the distinguishing capacity of the classifier. It means that the chosen features, set by the mentioned feature selection techniques provide the best combination of features given that improves the capability of a credit models to correctly identify the behavior of a potential borrower to pay back a loan.

After comparing and sorting the results of AUC measures for the classification algorithm, the procedure of selecting best one takes place. Besides, in order to achieve better performance, the ensemble of chosen models can be developed.

A hybrid approach for feature selection analysis

To present an argument for reasonable feature selection, it is advisable to focus not only on the meaning of the measures described above. We suggest to implement a hybrid approach.

We propose to average the values of the statistical measures for each model for different number of fields, which will allow to obtain the number of features for selection that corresponds to the highest accuracy of the model. Taking into account active use of Principal Component Analysis, presented in the literature on this issue, it is also advisable to take advantage of the following approach: the weight for the PCA will be 0.5, and the remainder will be evenly distributed among the statistical measures (0.125 for each).

Based on the hybridization of the measures, we can arrive at conclusion that the AUC values depend (or not) on the number of features to be selected, and then we will obtain the optimal number of features and develop model using the best chosen quality approach.

To recognize the correspondence between the selection criteria and the models used, the AUC values for each measure for different number of features for each model have to be averaged. If a particular criterion takes precedence for most models, it can be used for feature selection and developing the ensemble of models

4 Experimental Results

Data pre-processing

Our experiment starts with the process of data pre-processing. First of all, data cleaning was performed (removing missing values and irrelevant features from the database). Fields were screened based on the following criteria: maximum percentage of missing values; maximum percentage of records in a single category and maximum number of categories as a percentage of records (for categorical fields), minimum coefficient of variation and minimum standard deviation (for numeric fields).

While proceeding data we found out that there is a conflicting coding scheme in the database. Numerical attributes had two ways of representing integer separators from fractional: a comma (field “WRK_EXPERIENCE”) and a semicolon (all other numeric attributes), or another example: a date of birth field that has "-" or "." separators. There are also two data formats in the WRK_NROFEMPLOYEES field - categorical and date. Gender (“Female”, “female”) is also indicated by different formulations.

To prepare the data for modeling, we create a numeric field of the borrower's age, the flag field of the borrower's gender, where we reclassify the errors of entering the gender data. Also we create new flag fields for the presence of a partner, attitude towards the army and reclassify the occupational names field to six occupations to facilitate further modeling.

Another important moment is creating new features in order to adjust business logic to the process of feature selection. Thus, we added new feature interaction rules: the amount of income per family member, the amount of income per child, the amount of loan per term and the amount of income per payments. The practicality of this step is noted in many sources [21, 23].

The role is set to target for the field that indicates whether or not a given customer defaulted on the loan. The potential target fields were EVER_1_DPD, EVER_30_DPD, EVER_60_DPD, EVER_90_DPD. We have chosen EVER_30_DPD as the target, because such loan delinquencies are beneficial for the bank, because the borrowers also pay delay penalty in addition to the loan repayments.

After data pre-processing a dataset of 34 fields was prepared instead of initial 62. It consists of 41687 records. The data is in several formats: numeric (24), categorical (8) and logical (2).

It should be mentioned that the resulted dataset was split into two samples: training (75%) and testing (25%). To correct imbalances in dataset we use Balance node that causes an artificial increase records for which the target field EVER_30_DPD returned "1". The process of balancing data is essential in order to decrease misbalanced in initial data where the percentage of non-repayable loan is low in comparison with repayable ones (that is typical in credit-scoring problem for real business cases). Since many modeling techniques have trouble with biased data, they will tend to learn only the positive cases (repayable loans) and ignore the negative ones. If the data are well balanced with approximately equal numbers of positive and negative cases, models will have a better chance of finding patterns that distinguish the two groups. In this case, a Balance node is useful for creating a balancing directive that increases cases for non-repayable loans. In order not to misrepresent the true distribution results we apply Balance node only to the training sample of the data.

Feature Selection Techniques and Modeling

This section provides us with AUC values for scope of methods with and without feature selection including expert approach model (Table 1).

Table 1. AUC results for scope of models.

Model	AUC value
Logistic Regression	0.601
Neural Networks	0.596
SVM	0.583
CHAID	0.696
Expert	0.654

Without using feature selection, Decision Tree (CHAID) model has demonstrated the best results while pure expert model showed second separation power When scoring data uses Feature Selection node, the top n fields based on importance (4 different statistical measures) were selected (n = 30, 25, 20, 15, 10). Similar actions were performed for the PCA/Factor node. Next, Logistic Regression, Neural Network, Support Vector Machines, and Decision Tree (CHAID models were built for each case. The choice of these models is explained by the results of the application of Auto Classifier node.

According to expert approach, 20 factors were determined as significant. The most valuable features are ‘Loan payment to income ratio’, ‘Income to expenses ratio of borrower’, ‘Spouse availability’ and `Age of a borrower`. By significance all factors can be divided into 4 groups with same level of predictive strength with number of factors 3, 1, 10 and 6. By logical criteria all factors can be assigned to social-demographic, lending terms, financial state and lending history types. Further number of factors change is impractical as it is time consuming and contradicts to the aim of pure expert approach.

AUC results for models with feature selection for n equal to 25, 20 and 15 are presented in the Tables 2-5. Results for n equal to 10 and 30 are missed due to insignificant difference from 25 and 15 options respectively.

Table 2. AUC results due to feature selection technique for the Logistic Regression.

Number of features	(1) Pearson chi-square	(2) Likelihood-ratio chi-square	(3) Cramer's V	(4) Lambda	(5) PCA	(6) Average (1-4)	(7) Average (5-6)
15	0.740	0.749	0.746	0.745	0.681	0.745	0.713
20	0.748	0.749	0.747	0.749	0.594	0.748	0.671
25	0.746	0.748	0.747	0.747	0.589	0.747	0.668

Table 3. AUC results due to feature selection technique for the Neural Network.

Number of features	(1) Pearson chi-square	(2) Likelihood-ratio chi-square	(3) Cramer's V	(4) Lambda	(5) PCA	(6) Average (1-4)	(7) Average (5-6)
15	0.702	0.715	0.704	0.700	0.677	0.705	0.691
20	0.713	0.721	0.732	0.709	0.595	0.719	0.657
25	0.648	0.683	0.705	0.701	0.592	0.684	0.638

Table 4. AUC results due to feature selection technique for the SVM.

Number of features	(1) Pearson chi-square	(2) Likelihood-ratio chi-square	(3) Cramer's V	(4) Lambda	(5) PCA	(6) Average (1-4)	(7) Average (5-6)
15	0.657	0.673	0.67	0.678	0.507	0.670	0.588
20	0.679	0.671	0.672	0.679	0.582	0.675	0.629
25	0.638	0.673	0.686	0.685	0.580	0.671	0.625

Table 5. AUC results due to feature selection technique for the CHAID.

Number of features	(1) Pearson chi-square	(2) Likelihood-ratio chi-square	(3) Cramer's V	(4) Lambda	(5) PCA	(6) Average (1-4)	(7) Average (5-6)
15	0.702	0.675	0.688	0.716	0.686	0.695	0.691
20	0.695	0.684	0.718	0.69	0.697	0.697	0.697
25	0.729	0.668	0.714	0.721	0.680	0.708	0.694

It is obvious that PCA has proven to be an ineffective variable selection technique for CHAID and SVM, since the AUC values are close to the corresponding values in models without feature selection. The best results were achieved for logistic regression and model of neural networks according to criteria Likelihood-ratio chi-square and Cramer's V.

On average, the use of feature selection techniques improves the AUC value by 11.2% compared to the none-use. Note that the distribution of cumulative gain of AUC averages compared to the AUC averaged without feature selection is not uniform - the statistical criteria of Pearson chi-square, Likelihood-ratio chi-square, Cramer's V and Lambda improve by 13.0-14.7%, but Principal Component Analysis does only by 0.4%. That is why, it is advisable for the Ukrainian banks to try the alternative of feature selection by statistical measures, unlike the widespread foreign experience which prefers PCA [7, 11, 25].

The findings prove that in most cases with decreasing number of features, AUC measures for the classification algorithms increase, and cases reducing the variables to 20 improve models performance.

Hybrid approach of feature selection analysis

The tables 1-5 show the AUC results for a different number of input fields, selected in accordance with 5 feature selection criteria, as well as average values of 4 statistical measures and a weighted average of all presented measures (the maximum value is highlighted in grey).

AUC averages of all models (Table 6) confirm that the combination of statistical criteria allows to obtain the optimal number of features for modeling - 20, by PCA - 15.

Table 6. Average AUC results for the Logistic Regression, Neural Network, SVM, CHAID.

Number of features	(1) Pearson chi-square	(2) Likelihood-ratio chi-square	(3) Cramer's V	(4) Lambda	(5) PCA	(6) Average (1-4)	(7) Average (5-6)
15	0.700	0.703	0.702	0.710	0.638	0.704	0.671
20	0.709	0.706	0.717	0.707	0.617	0.710	0.663
25	0.690	0.693	0.713	0.714	0.610	0.702	0.656

It is obvious, that the value of AUC is higher using a uniform distribution of statistical measures, rather than using a weighted approach taking into account PCA.

An interesting fact is that with decreasing number of feature, we can observe their similarity tendency. The most frequently rated fields by feature selection technique are as follows:

1. CLN_YEARS (current age of a client);
2. WRK_FIELD (work field of a client);
3. SPOUSE (existence of a spouse);
4. INC_ALL_AP (total incomes of a client);
5. AMOUNT_PER_TERM (a sum of credit by term).

As it comes to the created features in order to adjust business logic to the process of feature selection, we should mention that all of them are selected for 20 important features by all criteria. Moreover, PCA takes them in the top six.

Attempts to determine the best feature selection technique by averaging AUC values in different models for different number of features were unsuccessful, as each model demonstrated different best measures): for the Logistic Regression - Likelihood-ratio chi-square, for Neural Network - Cramer's V, for SVM - Lambda, for CHAID - Pearson chi-square. As none of the criteria was overweight in most models, we concluded that the ensemble was inappropriate in this case.

5 Conclusion

The study shows processes of data pre-processing, feature creation and feature selection that can be applicable in real-life business situations for binary classification problems by using nodes from IBM SPSS Modeler. Results have proved that application of hybrid model of feature selection, which allows to obtain the optimal number of features, conduces in credit scoring accuracy increase.

Proposed hybrid model comparing to expert judgmental approach performs in harder explanation but shows better accuracy and flexibility of factors selection which is advantage in fast changing market.

Besides, the paper shows the accessibility of the approach for users of analytical platforms, the availability of tools in business application packages (IBM SPSS Modeler as an example) and the method of combining these tools. It is obvious that using feature selection techniques differ among countries, so the results of this particular study can be implemented in real business cases in Ukraine.

It should be noted that Ukrainian banks may be advised to try using the feature selection according to such statistical measures as Pearson chi-square, Likelihood-ratio chi-square, Cramer's V and Lambda, rather than PCA. The study results of Ukrainian lending market also show that the choice of features can be limit to 20, which allows to obtain the maximum AUC value.

The study empirically confirms that in Ukraine banks should consider hybrid selection technique with equal weights for statistical measures. The results show that the usage of a hybrid approach to feature selection methods improves the AUC value compared to the none-use by 11.2%, which is a clear advantage. On the other hand, the weak point of the approach is the increase of amount of time spent on calculations.

References

1. Abdou, H., Pointon, J.: Credit scoring, statistical techniques and evaluation criteria: a review of the literature. Intelligent Systems in Accounting, Finance & Management. 18 (2-3), 59-88 (2011) DOI: 10.1002/isaf.325
2. Aryuni, M., Madyatmadja, E.: Feature selection in credit scoring model for credit card applicants in xyz bank: A comparative study. International Journal of Multimedia and Ubiquitous Engineering. 10 (5), 17-24 (2015) DOI: 10.14257/ijmue.2015.10.5.03
3. Boughaci, D., Alkhawaldeh, A.A.: Three local search-based methods for feature selection in credit scoring. Vietnam Journal of Computer Science. 5, 107–121 (2018) DOI: 10.1007/s40595-018-0107-y
4. Chen, F.-L., Li, F.-C.: Combination of feature selection approaches with SVM in credit scoring. Expert Systems with Applications. 37 (7), 4902-4909 (2010) DOI: 10.1016/j.eswa.2009.12.025
5. Chornous, G., Nikolskyi, I.: Business-oriented feature selection for hybrid classification model of credit scoring. In.: Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), pp. 397-401. IEEE Press, Lviv (2018). DOI: 10.1109/DSMP.2018.8478534
6. Falangis, K., Glen, J.: Heuristics for feature selection in mathematical programming discriminant analysis models. Journal of the Operational Research Society. 61 (5), 804-812 (2010) DOI: 10.1057/jors.2009.24
7. Gietzen, T.: Credit Scoring vs. Expert Judgment - A Randomized Controlled Trial. SSRN Electronic Journal. (2017) DOI: 10.2139/ssrn.2983076.
8. Ha Van Sang, Ha Nam, Nguyen Duc Nhan: A Novel Credit Scoring Prediction Model based on Feature Selection Approach and Parallel Random Forest. Indian Journal of Science and Technology. 9(20), (2016) DOI: 10.17485/ijst/2016/v9i20/92299
9. IBM SPSS Modeler 18.2 Modeling Nodes. Copyright IBM Corp. 1994 – 2018. ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.2/en/ModelerModelingNodes.pdf
10. Kamalov, F., Thabtah, F.: A Feature Selection Method Based on Ranked Vector Scores of Features for Classification. Annals of Data Science. 4, 483–502 (2017). DOI: 10.1007/s40745-017-0116-1

11. Koutanaei, F., Sajedi, H., Khanbabaei, M.: A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. Journal of Retailing and Consumer Services. 27, 11-23 (2015) DOI: 10.1016/j.jretconser.2015.07.003
12. Kuhn, M., Johnson K.: Applied Predictive Modeling. 2nd ed. Springer (2018)
13. Kuhn, M., Johnson K.: Feature Engineering and Selection: A Practical Approach for Predictive Models. Chapman and Hall/CRC. (2019)
14. Liang, D., Tsai, C.-F., Wu, H.-T.: The effect of feature selection on financial distress prediction. Knowledge-Based Systems. 73 (1), 289-97 (2014) DOI: 10.1016/j.knosys.2014.10.010
15. Liu, Y., Schumann, M.: Data mining feature selection for credit scoring models. Journal of the Operational Research Society. 56 (9), 1099-1108 (2005) DOI: 10.1057/palgrave.jors.2601976
16. Louzada, F., Ara, A., Fernandes G.B. Classification methods applied to credit scoring: A systematic review and overall comparison. Surveys in Operations Research and Management Science. 21(2), 117-134 (2016) DOI: 10.1016/j.sorms.2016.10.001
17. Rozlini, M., Munirah, M. Y., Wahidi, N.: A Comparative Study of Feature Selection Techniques for Bat Algorithm in Various Applications. MATEC Web of Conferences, 150 (2018) 06006 DOI: 10.1051/matecconf/201815006006
18. Sadatrasoul, S., Gholamian, M., Shahanaghi, K.: Combination of feature selection and optimized fuzzy apriori rules: The case of credit scoring. International Arab Journal of Information Technology. 12 (2), 138-145 (2015)
19. Salappa, A., Doumpos, M., Zopounidis, C.: Feature selection algorithms in classification problems: an experimental evaluation. Optim. Methods Softw. 22(1), 199–212 (2007) DOI: 10.1080/10556780600881910
20. Shi, J., Zhang, S.-Y., Qiu, L.-M.: Credit scoring by feature-weighted support vector machines. Journal of Zhejiang University: Science C. 14 (3), 197-204 (2013) DOI: 10.1631/jzus.C1200205
21. Siddiqi, N.: Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards. 2 ed. Wiley (2017)
22. Somol, P., Baesens, B., Pudil, P., Vanthienen, J.: Filter- versus wrapper-based feature selection for credit scoring. International Journal of Intelligent Systems. 20 (10), 985-999 (2005) DOI: 10.1002/int.20103
23. Thomas, L. C., Edelman D.B., Crook J.N.: Credit scoring and its applications. SIAM-Society for Industrial & Applied Mathematics. 2nd revised ed. (2017)
24. Tripathi, D., Edla, D., Kuppili, V., Bablani, A., Dharavath, R. Credit Scoring Model based on Weighted Voting and Cluster based Feature Selection. Procedia Computer Science. 132, 22-31 (2018). DOI: 10.1016/j.procs.2018.05.055
25. Van-Sang Ha, Ha-Nam Nguyen: Credit scoring with a feature selection approach based deep learning. MATEC Web of Conferences, 54 (2016) DOI: 10.1051/matecconf/20165405004
26. Volkova, E.S., Gisin, V.B., Solov'ev, V.I.: Data Mining Techniques: Modern Approaches to Application in Credit Scoring. Finance and Credit. 23(34), 2044–2060 (2017) DOI: 10.24891/fc.23.34.2044
27. Waad, B., Ghazi, B., Mohamed, L.: A three-stage feature selection using quadratic programming for credit scoring. Applied Artificial Intelligence. 27 (8), 721-742 (2013) DOI: 10.1080/08839514.2013.823327
28. Wang, J., Hedar, A.-R., Wang, S., Ma, J.: Rough set and scatter search metaheuristic based feature selection for credit scoring. Expert Systems with Applications. 39 (6), 6123-6128 (2012) DOI: 10.1016/j.eswa.2011.11.011