Bank i Kredyt 54(4), 2023, 419-436

BANK I KREDYT
BANK & CREDIT

Comparison of different approaches using Random Forest for imbalanced credit data

Anna Matuszyk*

Submitted: 6 March 2023. Accepted: 31 June 2023.

Abstract

Credit scoring models are extensively used in credit risk management of individual customers. These models are based on econometric methods using past data about customers, both defaulters and non--defaulters. These models focus on the optimal separation between good and bad customers taking into account two types of errors that appear, namely: the False Positive (Type 1 error) and the False Negative (Type 2 error).

The purpose of the project was to focus on the problem of unbalanced data. Different balancing methods have been applied to the data set obtained from the financial institution operating in the European market. Various levels of unbalance have been considered and different statistical assessment metrics have been compared.

Keywords: credit scoring models, unbalanced data, balancing technique, Random Forest, model performance

JEL: C01, C13, C52

* Warsaw School of Economics, Collegium of Management and Finance, Financial System Department; e-mail: amatus1@sgh.waw.pl; ORCID: 0000-0002-9414-2784.

1. Introduction

Nowadays, the amount of data is growing constantly, both in size and in complexity. This phenomenon is also present in banking, where information about customers is being gathered. One of the issues is a classification problem, especially when different classes are not balanced. The number of imbalanced classification methods increased, but in the majority of cases they focus on normal-sized data sets. In this project, the analysis will be conducted in the context of imbalanced data, using different resampling methods.

The unbalanced data present one of the most interesting and important problems in data mining. An additional challenge that coexists in banking, especially in credit risk, is a classification problem, especially when different classes are not balanced. It is worth checking this issue and selecting the most appropriate balancing methods for specific problems.

The main objective of this project is to evaluate different balancing techniques for the credit data sets. Unbalanced classes in the data sets pose a challenge during the classification process. The majority of the research done so far is focused mainly on dealing with the ratio of the unbalanced sample and does not consider other problems associated with such data. One of them is the inclusion of type 1 and type 2 errors, which are contented with the classification. High measures of the model performance are not sufficient criteria that decide which model to use. It may turn out that despite the satisfactory value of AUC, the error value eliminates the model from usage.

2. Literature review

Unbalanced data sets are challenging to analyse. The main reason is the fact that the algorithms applied to solve the problem do not cope with the number of observations between two classes. The unbalanced nature of the data is typical for credit data sets, where the number of defaulted customers is much smaller than the defaulted ones.

According to the Basel Committee on Banking Supervision (BIS 2005), the calculations based on historical data made for very safe assets may “not be sufficiently reliable” for estimating the probability of default. The reason for this is that because there are so few defaulted observations, the resulting estimations are likely to be inaccurate. Therefore there is a need for a better understanding of the appropriate modelling techniques for data sets which display a limited number of defaulted observations.

In the literature many different approaches were proposed to deal with the problem of the unbalanced data. One of them relies on assigning a higher cost for the misclassification. This was tried by Domingos (1999) and Pazzani et al. (1994). The problem of imbalanced data was studied by Shi et al. (2023), who applied a hybrid classification model based on data density.

Research conducted by Niu et al. (2020) focused on misclassification of the loan applicants. According to the authors, class imbalance of data is a factor that affects the classification performance of the model. This encouraged them to use a novel ensemble model based on data distribution for an imbalanced credit risk sample. The results obtained show that this approach not only delivered a good performance, but also improved the classification performance.

An interesting and wide study was performed by Baesens et al. (2003). In this benchmarking study, the authors compared seventeen techniques on eight credit data sets. The performance was assessed using the classification accuracy and the area under the receiver operating characteristic curve (AUC). This research was extended and several novel classification algorithms were compared by Lessmann et al. (2015). The authors compared 41 classification methods across eight credit scoring data sets. The results obtained suggested that several classifiers achieved significantly more accurate predictions than the standard logistic regression.

Another comparison of different data mining techniques was made by Yeh and Lien (2009). It was found that the artificial neural networks model achieved the highest values of R-square in estimating the real probability of default. This finding was in line with Baesens et al. (2003).

Another approach, proposed by Japkowicz (2000) was based on under- and over-sampling techniques and compared with her own method called "learning by recognition". According to her findings, over--sampling as well as under-sampling can be very effective methods. This approach was extended by Chawla et al. (2002). The authors proposed the usage of the Synthetic Minority Over-sampling Technique – SMOTE. The results obtained showed that such an approach can improve the accuracy of classifiers for a minority class. Zhu et al. (2019) applied the Random Forest (RF) algorithm in order to build a prediction model for the loan data set. The authors used the SMOTE method to solve the problem of imbalance class in the data set. According to the results obtained, the Random Forest algorithm outperformed other methods, namely: logistic regression, decision tree and other machine learning algorithms.

A similar approach was undertaken by Abedin, Guotai and Hajek (2022), who also applied the extended SMOTE technique to overcome the problem with imbalanced credit data. The results from the study proved that the applied sampling method outperformed non-sampling algorithms. Moreover, Random Forest turned out to be a good choice for the target modelled.

Weiss and Provost (2003) tried to find out which good/bad distribution is the most appropriate in classifying a data set. It was found that the optimal class distribution should contain between 50% and 90% minority class examples within the training set. Similarly, Namvar et al. (2018) compared different combinations of classifiers and resampling techniques using the imbalanced data. According to their findings, combining Random Forest and random under-sampling can be an effective strategy in order to calculate the credit risk in social lending markets. Rao et al. (2020) applied sensitive Random Forest model to evaluate the credit risk of the borrowers.

Another approach was proposed by Mqadi, Naicker and Adeliy (2021), who applied a novel technique to cope with the problem of imbalanced data sets. The authors proposed using Random Forest and a hybrid data-point approach. Achieved results were compared with the results of logistic regression, support vector machine, decision tree, and Random Forest. The proposed approach improved the predictive accuracy of all the algorithms tried with the dominant advantage of Random Forest.

A very important part of the credit risk modelling is the analysis of the misclassification cost, as it directly affects the profitability of the creditor. To properly minimize the cost of misclassification, lenders should make a careful analysis. This analysis includes consideration of different thresholds that determine whether an applicant will be granted a credit or will be rejected. The two errors associated with this should also be considered, namely: accepting a bad customer and rejecting a good one. The misclassification cost is an important part of the credit risk modelling process and has been studied in the literature. Bahnsen, Aouada and Ottersten (2015) proposed a cost-sensitive decision tree algorithm. The authors applied different example-dependent costs into a new cost-based impurity measure and a new cost-based pruning criteria. Using different data sets, the proposed approach was used and evaluated. According to the results obtained, the proposed algorithm gave promising results.

The answer to the problem of customers' misclassification is also more recent scientific studies that combine cost-sensitive learning (CSL) conducted by: Shen, Wang, Shen (2019), Xia, Liu, Liu (2017), Xiao et al. (2020). If credit scoring models improperly classify risk-free borrowers and therefore reject their loan applications, financial institutions face only opportunity costs. On the contrary, financial institutions are likely to suffer significant losses if they accept risky borrowers. Overall, the studies suggest that misclassification costs are a significant factor in credit modelling, and that careful analysis of misclassification costs is necessary to get accurate and cost-effective credit risk models.

Although many different studies have been conducted considering imbalanced credit data, there is still a potential for more detailed work to be conducted. According to the author's best knowledge, there is no study considering the influence of balancing technique together with misclassification costs, and the different number of trees in Random Forest. There is also a lack of studies performed for Poland, so this paper fulfils this gap as the data comes from one of the financial institutions operating on the Polish market.

3. Two types of errors

Credit scoring is a tool used to analyse the borrower's risk. It is a mathematical and statistical instrument used to assess creditworthiness. Scoring models classify clients according to the degree of the risk associated with them. Models provide an objective assessment of creditworthiness carried out according to the same criteria for all clients.

Two kinds of errors can arise when building such models, which are related to two kinds of costs. The first one classifies a good customer into the bad group, and therefore that person is rejected. In this case, the potential profit from this applicant is a loss. The second type of error may arise when a bad client is classified into the group of the good ones. In this case, the loss appears when the customer stops paying off their obligation.

The control of the scoring model makes it possible to determine the level of type 1 and type 2 errors when classifying the customers. These errors are defined as follows:

Type 1 – rejection of an applicant that should be approved,

Type 2 – approval of an applicant that should be rejected.

The first type of error is related to rejecting the credit to a customer that fails to fulfil its obligations. The second type of error is related to granting the credit to a client that should be rejected. In this situation, the model builder has to find the best balance between type 1 and type 2 error.

4. Data, balancing techniques, method and measures used in the analysis

4.1. Data set description

The data set used in this research study is a portfolio of the leases granted for the customers, and coming from a bank (which prefers to stay anonymous) operating on the Polish market. The total data set consists of 10,993 cases, including 350 defaults, provided on the customer level. It was split into a training and a test sample, so the training sample contains 7,695 cases, including 252 default ones.

The test sample contains 3,298 cases, including 98 defaults. The product type is leasing. Table 1 summarizes the training sample, which is used for the model estimation; and the test sample, which is reserved for assessing the model's predictive accuracy. Splitting the data into training and test samples is a common technique in credit scoring. For the purpose of this analysis the split is 70% : 30%. The balancing techniques were applied only to the training sample.

Since it would be good to see an analysis of the robustness to different splits between the training and the test sets, an additional (shorter) analysis was done for a different split of the raw data. The original sample was split so that the training sample consisted of 8,754 cases, including 210 defaults, and the test sample consisted of 2,239 observations, including 140 defaults. The results for the test sample are presented in Table 4.

The lease agreement ranged from 12 to 72 months. The lease was offered for small and medium enterprises. The data contains information about customers and lease, namely: status (good or bad), branch and age of the company, location, car type and age, amount of the lease and monthly instalments, etc.

4.2. Application of the balancing techniques

In this research three main balancing techniques were used, namely:

- Under-sampling (where good customers in the training sample were removed):
  - the main proportion 1:1,
  - additional proportion 2:1;
- Over-sampling (where bad customers in the training sample were replicated):
  - the main proportion 1:1,
  - additional proportion 1:2;
- Both: under- and over-sampling:
  - the main proportion number of observations: 1000,
  - the additional proportion number of observations: 1500.

In order to determine the optimal ratio of under- and over-sampling, many trials were taken with different proportions and error costs. The benchmark original imbalanced training sample was used in order to check whether the techniques applied affect the prediction. These allowed a comparison of all the results obtained (Table 2).

It is important to mention that under- and over-sampling were performed only for the training sample, not for the test sample. The performance measurement has been received by different balancing techniques and compared with the full training data set. The test sample remained unchanged in order to provide unbiased results of the model performance.

According to Japkowicz and Stephen (2002), the problem of unbalanced data is dependent on four factors:

– the degree of class imbalance,
– the complexity of the concept represented by the data,
– the overall size of the training data,
– the type of the classifier.

Alberto et al. (2018) suggest that the degree of class imbalance can provide information about the data imbalance and can help structure the strategy for dealing with it.

As a performance measure, the AUC (area under the curve) and type 1 and type 2 errors were chosen. The higher the AUC value, the better the performance of the scoring model. The AUC takes the values from 0 to 1. The receiver operating characteristic curve (usually called ROC) is a two-dimensional graph presenting the relation between the true positive rate (sensitivity) and the false positive rate (1-specificity). In order to compare the ROC curves of different models built, the area under the receiver operating characteristic curve (AUC) is calculated. An example of the ROC curve is presented in Figure 1. The diagonal line represents a random scoring model where sensitivity = 1-specificity. According to this, good classification is when the ROC curve is about the diagonal line and AUC is greater than 50%.

In the credit scoring context, sensitivity is interpreted as a cumulative proportion of defaults above a certain score *s* (correctly rejected) and 1-specificity – as a cumulative proportion of non-defaults incorrectly rejected (Thomas, Edelman, Crook 2002). The higher values of AUC suggest more superior models, no discrimination corresponds to AUC with value 0.5.

4.3. Misclassification costs

The purpose of the scoring model is to assign a customer to a group of good or bad customers, but, as mentioned earlier, two types of errors can appear, error type 1 and error type 2. In order to minimize these errors, misclassification costs were applied. In the next step, the results of applying different misclassification costs to the built models were analysed. The comparison was based on all models built in the former step, and carried out for misclassification costs as follows:

Mc1 = (1,2)
Mc2 = (1,3)
Mc3 = (1,4)

where:

- value 1 – was assigned for misclassification of good customers to the bad ones (error type 1);
- values 2, 3, 4 – were assigned for misclassification of bad customers to the group of the good ones (error type 2).

4.4. Random Forest

Random Forest is a machine learning method of classification, trained on bootstrap samples of the training data using random feature selection in the process of tree generation. Random Forest is becoming a more and more popular technique as it avoids problems associated with a single classification tree, such as instability of the trees (high sensitivity to small changes in the sample), the risk of "overfitting" and the need of pruning the tree. There are two parameters that need to be considered for the Random Forest, namely: the number of trees and the number of attributes used to grow each tree. A more detailed explanation of how to train a Random Forest can be found in Breiman (2001).

Random Forest is a popular machine learning algorithm used for classification, regression and other tasks. This ensemble method combines multiple decision trees to improve predictive accuracy and reduce over-fitting. Random Forest is characterized by hyperparameters that can be adjusted to optimize the model's performance. Table 5 presents the most important ones considered in the research.

5. Results

The results (Table 3 and Table 4) from this empirical study indicate that all classification approaches perform well in the case of AUC values. Over-sampling (without considering misclassification costs) performs significantly better than the under-sampling or when using both classifiers. However, considering the results of the model with the values of type 1 and type 2 error becomes more challenging as the high values of AUC do not correspond with the low values of these errors.

Adding misclassification costs didn't affect the results as much as the number of trees in RF. It is visible that models where 1000 trees were used performed better than those with 200 trees.

Generally speaking, it can be summarized that:

– over-sampling technique performs the best for the data set used, both with and without considering misclassification costs;

– number of trees in RF influences models' results (models with 200 trees performed slightly worse than those with 1000 trees);

– misclassification costs didn't affect results as it was initially assumed;

– AUC cannot be the only measure considered when choosing the most appropriate model; although for some models the AUC value was very high, the errors (total, type 1 and type 2) didn't look reasonable;

– under-sampling technique performed the worst in the case of the total error: but when considering type 1 and type 2 errors, it turned out that type 2 error was the lowest when using this balancing technique, especially when the sampling proportion 1:1 was applied;

– over-sampling performed the best in the case of AUC values and type 1 errors (the lowest); unfortunately, for type 2 errors it received high values, which means that bad customers were classified as good ones. This means a loss for a financial institution.

The final decision belongs to the institution, which would decide what compromise between AUC, type 1 and type 2 errors could be accepted. That is why focusing only on AUC values is not the proper approach and can be misleading. The proper analysis should include information about errors when choosing the final model.

6. Summary

The paper explains how AUC along with type 1 and type 2 errors play a crucial role in building scoring models. In this case the results of the analysis were shown and explained.

In addition to standard measures of model fit, credit scoring models are evaluated in terms of their ability to discriminate between 'good' and 'bad' credit risk. Area under the curve (AUC) is a common measure for the discriminatory power.

In this study, data balancing techniques were analysed and their performance was studied over various aspects additionally used as error costs and number of trees. The classification power of the models built was assessed based on the area under the receiver operating characteristic curve (AUC). These were compared with errors (type 1 and type 2). The misclassification cost analysis is an important factor when building accurate credit risk models. The cost of misclassification reflects the potential financial loss that a lender faces in the case of misclassifying a borrower's credit risk. This has been confirmed not only by the scientific studies cited in this paper, but also by the analysis results presented. Therefore, this part of the model building process is very important and should be taken into account, for example, by means of different weights assigned to errors of the first and second type, so as to choose the most optimal set.

In the future, some further analysis will be conducted in order to investigate other approaches to the imbalanced credit data. In this case, more data sets will be acquired with larger data volume (more observations), which will allow the inclusion of other methods.

References

Abedin M.Z., Guotai C., Hajek P. (2022), Combining weighted SMOT10.1007E with ensemble learning for the class-imbalanced prediction of small business credit risk, *Complex and Intelligent Systems,* DOI: 10.1007/s40747-021-00614-4.

Alberto F., García S., Galar M., Prati R., Krawczyk B., Herrera F. (2018), *Learning from Imbalanced Data Sets*, Springer Nature Switzerland AG.

Baesens B., Van Gestel T., Viaene S., Stepanova M., Suykens J., Vanthienen J. (2003), Benchmarking state of the art classification algorithms for credit scoring, *Journal of the Operational Research Society*, 54(6), 627–635.

Bahnsen A.C., Aouada D., Ottersten B. (2015), Example-dependent cost-sensitive decision trees, *Expert Systems with Applications*, 42(19), 6609–6619.

BIS (2005), *Basel committee newsletter no. 6: Validation of low-default portfolios in the Basel II framework. Technical report*, Bank for International Settlements, Basel Committee on Banking Supervision.

Breiman L. (2001), Random Forests, *Machine Learning*, 45(1), 5–32.

Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. (2002), SMOTE: Synthetic Minority Over--sampling Technique, *Journal of Artificial Intelligence Research*, 16, 321–357.

Domingos P. (1999)*, MetaCost: a general method for making classifiers cost-sensitive*, KDD, 99: Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining, DOI: 10.1145/312129.312220.

Japkowicz N. (2000), *Learning from imbalanced data sets: a comparison of various strategies*, https://sci2s.ugr.es/keel/pdf/specific/congreso/aaai2000-workshop.pdf.

Japkowicz N., Stephen S. (2002), The class imbalance problem: a systematic study, *Intelligent Data Analysis*, 6, 429–449.

Lessmann S., Baesens B., Seow H.-V., Thomas L.C. (2015), Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research, *European Journal of Operational Research*, 247(1), 124–136.

Mqadi N.M., Naicker N., Adeliy T. (2021), Solving misclassification of the credit card imbalance problem using near miss, *Mathematical Problems in Engineering*, DOI: 10.1155/2021/7194728.

Namvar A., Siami M., Rabhi F., Naderpour M. (2018), Credit risk prediction in an imbalanced social lending environment, *Computer Science*, https://arxiv.org/abs/1805.00801.

Niu K., Zhang Z., Liu Y., Li R. (2020), Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending, *Information Sciences*, 536, 120–134.

Pazzani M.J., Merz C., Murphy P., Ali K., Hume T., Brunk C. (1994), Reducing misclassification costs, in: *Proceedings of the Eleventh International Conference on Machine Learning*, Morgan Kaufmann.

Rao C., Liu M., Goh M., Wen J. (2020), A 2-stage modified random forest model for credit risk assessment of P2P network lending to "Three Rurals" borrowers, *Applied Soft Computing*, 95.

Shi S., Li J., Zhu D., Yang F., Xu Y. (2023), A hybrid imbalanced classification model based on data density, *Information Sciences*, 624, 50–67.

Shen F., Wang R., Shen Y. (2019), A cost-sensitive logistic regression credit scoring model based on multi-objective optimization approach, *Technological and Economic Development of Economy*, 1–25, DOI: /10.3846/tede.2019.11337.

Thomas L.C., Edelman D.B., Crook J.N. (2002), *Credit Scoring and Its Applications*, SIAM.

Weiss G., Provost F. (2003), Learning when training data are costly: the effect of class distribution on tree induction, *Journal of Artificial Intelligence Research*, 19, 315–354.

Yeh I.C., Lien C.H. (2009), The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, *Expert Systems with Applications*, 36(2), 2473–2480.

Xia Y., Liu C., Liu N. (2017), Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending, *Electronic Commerce Research and Applications*, 24, 30–49, DOI: 10.1016/j.elerap.2017.06.004.

Xiao J., Zhou X., Zhong Y., Xie L., Gu X., Liu D. (2020), Cost-sensitive semi-supervised selective ensemble model for customer credit scoring, *Knowledge-Based Systems*, 189, 105118, DOI: 10.1016/j.knosys.2019.105118.

ZhuL., Qiua D., Ergua D., Yinga C., Liu K. (2019), A study on predicting loan default based on the random forest algorithm, *Procedia Computer Science*, 162, 503–513.

Appendix

Table 1

Training and test samples

	Training sample			Test sample		
	good	bad	total	good	bad	total
Number of customers	7 443	252	7 695	3 200	98	3 298

Source: own calculation.

Table 2

Sampling proportions

Sampling approach	Sampling ratio/size of the sample	Number of good customers	Number of bad customers
Undersampling	1:1	252	252
	1:2	504	252
Oversampling	1:1	7 443	7 443
	2:1	7 443	14 886
Both (under- and oversampling)	1 000	491	509

Source: own calculation.

Table 3
Models' results

Type of sampling	Train sample 7 443 good/252 bad							Test sample 3 200 good/98 bad					
	Proportion (G:B)	No. of trees	Error (%)	G as G	Type 1 error G as B	Type 2 error B as G	B as B	Accuracy	G as G	Type 1 error G as B	Type 2 error B as G	B as B	Error (%)
Imbalanced	7443G/252B	200	2.43	7440	3	184	68	0.9773	3197	3	72	26	2.27
Imbalanced	7443G/252B	1000	2.43	7439	4	183	69	0.9776	3198	2	72	26	2.24
Under sampling 1:1	252G/252B	200	20	194	58	43	209	0.822	2623	577	10	88	17.8
Under sampling 1:1	252G/252B	1000	17.7	201	51	38	214	0.7962	2536	664	8	90	20.4
Under sampling 1:1 (costs 1,2)	252G/252B	200	17.1	206	46	40	212	0.7929	2527	673	10	88	20.7
Under sampling 1:1 (costs 1,3)	252G/252B	200	0.19	199	53	42	210	0.7878	2511	689	11	87	21.2
Under sampling 1:1 (costs 1,4)	252G/252B	200	21.4	197	55	53	199	0.8235	2629	571	11	87	17.7
Under sampling 1:1 (costs 1,2)	252G/252B	1000	0.18	200	52	38	214	0.8102	2587	613	13	85	19
Under sampling 1:1 (costs 1,3)	252G/252B	1000	21	191	61	45	207	0.8141	2596	604	9	89	18.6
Under sampling 1:1 (costs 1,4)	252G/252B	1000	17.3	201	51	36	216	0.8093	2579	621	8	90	19.1
Under sampling 1:2	504G/252B	200	16	457	47	74	178	0.9002	2887	313	16	82	9.98

Table 3, cont'd

Type of sampling	Train sample 7 443 good/252 bad							Test sample 3 200 good/98 bad					
	Proportion (G:B)	No. of trees	Error (%)	G as G	Type 1 error G as B	Type 2 error B as G	B as B	Accuracy	G as G	Type 1 error G as B	Type 2 error B as G	B as B	Error (%)
Under sampling 1:2	504G/252B	1000	16.8	456	48	79	173	0.9169	2945	255	19	79	8.31
Under sampling 1:2 (costs 1,2)	504G/252B	200	16.7	457	47	79	173	0.9066	2912	288	20	78	9.34
Under sampling 1:2 (costs 1,2)	504G/252B	1000	17.1	459	45	84	168	0.916	2944	256	21	77	8.4
Under sampling 1:2 (costs 1,3)	504G/252B	200	16	464	40	81	171	0.9118	2930	270	21	77	8.82
Under sampling 1:2 (costs 1,3)	504G/252B	1000	15.6	463	41	77	175	0.9166	2945	255	20	79	8.34
Under sampling 1:2 (costs 1,4)	504G/252B	200	16.5	459	45	80	172	0.9099	2921	279	18	80	9.01
Under sampling 1:2 (costs 1,4)	504G/252B	1000	15.5	464	40	77	175	0.8975	2883	317	21	77	10.3
Over sampling 1:1	7443G/7443B	200	0.1	7428	15	0	7443	0.9809	3194	6	57	41	1.91
Over sampling 1:1	7443G/7443B	1000	0	7435	8	0	7443	0.9809	3194	6	57	41	1.91
Over sampling 1:1 (costs 1,2)	7443G/7443B	200	0.1	7428	15	0	7443	0.9806	3193	7	57	41	1.94

Table 3, cont'd

Type of sampling	Train sample 7 443 good/252 bad							Test sample 3 200 good/98 bad					
	Proportion (G:B)	No. of trees	Error (%)	G as G	Type 1 error G as B	Type 2 error B as G	B as B	Accuracy	G as G	Type 1 error G as B	Type 2 error B as G	B as B	Error (%)
Over sampling 1:1 (costs 1,3)	7443G/7443B	200	0.09	7430	13	0	7443	0.9809	3192	8	55	43	1.91
Over sampling 1:1 (costs 1,4)	7443G/7443B	200	0.06	7434	9	0	7443	0.9806	3192	8	56	42	1.94
Over sampling 1:1 (costs 1,2)	7443G/7443B	1000	0.07	7433	10	0	7443	0.98	3192	8	58	40	2
Over sampling 1:1 (costs 1,3)	7443G/7443B	1000	0.07	7433	10	0	7443	0.9797	3192	8	59	39	2.03
Over sampling 1:1 (costs 1,4)	7443G/7443B	1000	0.07	7433	10	0	7443	0.9803	3192	8	57	41	1.97
Over sampling 1:2	7443G/14886B	200	0.07	7428	15	0	14886	0.9803	3192	8	57	41	1.97
Over sampling 1:2	7443G/14886B	1000	0.05	7431	12	0	14886	0.9809	3192	8	55	43	1.91
Over sampling 1:2 (costs 1,2)	7443G/14886B	200	0.05	7432	11	0	14886	0.98	3191	9	57	41	2
Over sampling 1:2 (costs 1,2)	7443G/14886B	1000	0.05	7431	12	0	14886	0.9806	3192	8	56	42	1.94
Over sampling 1:2 (costs 1,3)	7443G/14886B	200	0.07	7428	15	0	14886	0.98	3189	11	55	43	2

Table 3, cont'd

Type of sampling	Train sample 7 443 good/252 bad							Test sample 3 200 good/98 bad					
	Proportion (G:B)	No. of trees	Error (%)	G as G	Type 1 error G as B	Type 2 error B as G	B as B	Accuracy	G as G	Type 1 error G as B	Type 2 error B as G	B as B	Error (%)
Over sampling 1:2 (costs 1,3)	7443G/14886B	1000	0.04	7433	10	0	14886	0.9803	3191	9	56	42	1.97
Over sampling 1:2 (costs 1,4)	7443G/14886B	200	0.08	7426	17	0	14886	0.98	3192	8	58	40	2
Over sampling 1:2 (costs 1,4)	7443G/14886B	1000	0.04	7433	10	0	14886	0.9797	3191	9	58	40	2.03
Both 1000	491G/509B	200	6.4	440	51	13	496	0.8993	2889	311	21	77	10.1
Both 1000	491G/509B	1000	6.6	436	55	11	498	0.8987	2888	312	22	76	10.1
Both 1000 (costs 1,2)	491G/509B	200	6.4	439	52	12	497	0.9015	2895	305	20	78	9.85
Both 1000 (costs 1,3)	491G/509B	200	6.1	443	48	14	496	0.8933	2867	333	19	79	10.7
Both 1000 (costs 1,4)	491G/509B	200	6.2	442	49	13	496	0.8945	2871	329	19	79	10.6
Both 1000 (costs 1,2)	491G/509B	1000	6.6	438	53	13	496	0.8972	2880	320	19	79	10.3
Both 1000 (costs 1,3)	491G/509B	1000	6.7	437	54	13	496	0.8963	2878	322	20	78	10.4
Both 1000 (costs 1,4)	491G/509B	1000	6.2	440	51	11	498	0.899	2886	314	19	79	10.1

Source: own calculation.

Table 4
Additional models' results (for the test sample only)

Type of sampling	Proportion (G:B) in the train sample 8544:210	AUC on test (no cost)	AUC cost 1,2	AUC cost 1,3	AUC cost 1,4	AUC cost 1,5	AUC cost 2,3	AUC cost 2,5	AUC cost 3,5
Imbalanced sample	8544/210 defaults	0.677	0.769	0.732	0.752	0.751	0.764	0.769	0.759
Over balanced 1:1	8544/8544 defaults	0.809	0.805	0.796	0.811	0.811	0.82	0.804	0.82
Over balanced 1:2	8544/17088 defaults	0.808	0.823	0.823	0.823	0.807	0.796	0.823	0.793
Over balanced 1:3	8544/25632 defaults	0.817	0.812	0.817	0.817	0.817	0.812	0.815	0.812
Over balanced 1:4	8544/35772 defaults	0.863	0.819	0.813	0.813	0.815	0.819	0.813	0.819
Under balanced 1:1	210 censored/210 defaults	0.828	0.801	0.794	0.793	0.793	0.802	0.794	0.802
Under balanced 1:2	420 censored/210 defaults	0.821	0.822	0.821	0.819	0.815	0.834	0.822	0.834
Under balanced 1:3	630 censored/210 defaults	0.819	0.8	0.816	0.813	0.807	0.801	0.817	0.804
Under balanced 1:4	840 censored/210 defaults	0.818	0.781	0.803	0.829	0.826	0.769	0.763	0.791
Both 500	270 censored/230 defaults	0.764	0.759	0.798	0.766	0.766	0.762	0.761	0.762
Both 900	466 censored/434 defaults	0.824	0.822	0.828	0.826	0.83	0.832	0.826	0.832
Both 1000	520 censored/480 defaults	0.844	0.795	0.802	0.802	0.792	0.798	0.795	0.798
Both 2000	1039 censored/961 defaults	0.76	0.811	0.799	0.802	0.802	0.817	0.811	0.808
Both 1500	789 censored/711 defaults	0.817	0.842	0.817	0.82	0.833	0.826	0.85	0.842

Number of trees = 1000.
Test sample proportion: 2099G/140B.

Source: own calculation.

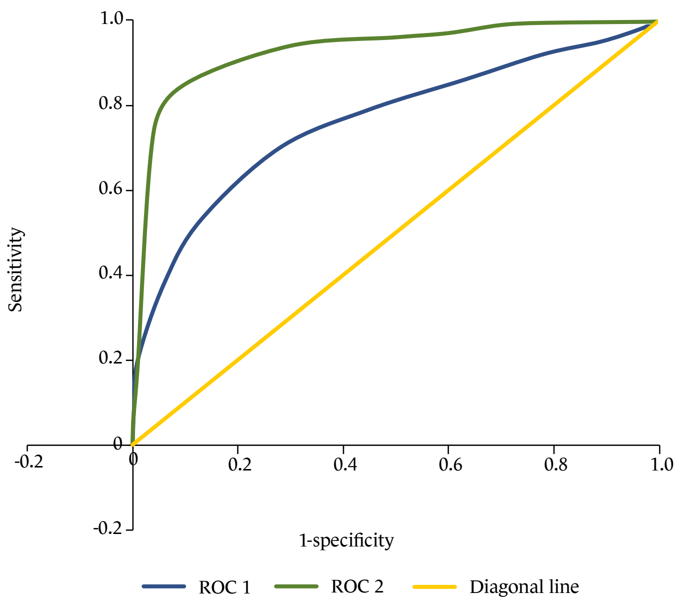
Table 5

Hyperparameters for the Random Forest

Name of the hyperparameter	Description	Details of the parameter used in the analysis
Package name/software	N/A	Random Forest in R
Number of trees	Number of decision trees used in the forest	200 and 1000
Type of Random Forest	Purpose of the model	classification
Tree depth	Maximum depth of each decision tree in the forest	6
Number of features	Number of features to consider at each split of the decision tree	6
Sample size	Number of cases to be used for training each decision tree in the forest	different for each balanced sample
Forest terminal node size	Minimum size of terminal nodes. Setting this number larger causes smaller trees to be grown	10
Total number of variables	Total number of variables used as an input	37
Split criterion	Criterion used for splitting the decision tree	logrank

Source: own elaboration based on the literature.

Figure 1
Example of different ROC curves



Porównanie różnych podejść wykorzystujących lasy losowe dla niezbilansowanych danych kredytowych

Streszczenie

Tempo rozwoju zaawansowanych technologii z roku na rok staje się coraz szybsze. Ilość gromadzonych danych stale rośnie. Jednocześnie możliwe jest prowadzenie badań związanych z posiadanymi danymi. Jedną z kwestii, na którą warto zwrócić uwagę, jest problem danych niezbilansowanych. Ten typ danych charakteryzuje się znaczną dysproporcją między przypadkami reprezentującymi poszczególne klasy. Liczba obserwacji należących do analizowanej klasy (nazywanej klasą mniejszościową) jest znacznie mniejsza niż liczba pozostałych obserwacji (nazywanych klasą większościową). Przedmiotem zainteresowania w niniejszym projekcie badawczym będzie klasa, która w zbiorze danych ma niewystarczającą liczbę obserwacji.

Niezbilansowane dane są obecne w wielu różnych dziedzinach nauki, począwszy od badań nad trzęsieniami ziemi, pandemią, aż po kryzysy finansowe. W większości przypadków badacze są zainteresowani przewidywaniem wydarzeń z klasy mniejszościowej. W tym projekcie także przewidywano klasę mniejszościową; stanowili ją klienci, którzy zaprzestali spłaty kredytu. Wynika to z faktu, że generuje ona więcej problemów.

W tym celu rozważono różne poziomy niezbilansowania danych kredytowych oraz dodatkowe trudności wpływające na ocenę uzyskanych wyników. Ponadto zostały porównane różne metody bilansowania, a także wyniki modeli z różną liczbą drzew; uwzględniono też koszty złej klasyfikacji. Analizowany zbiór danych pozyskano od instytucji finansowej działającej na polskim rynku.

Na postawie otrzymanych wyników można stwierdzić, że nie ma optymalnego podejścia, które byłoby odpowiednie do rozwiązania wszystkich problemów występujących w tego rodzaju bazach danych.

W przyszłości zostaną przeprowadzone dalsze analizy w celu zbadania innych podejść do niezbilansowanych danych kredytowych. W tym celu zostaną pozyskane i przeanalizowane nowe zestawy danych.

Słowa kluczowe: modele scoringu kredytowego, dane niezbilansowane, technika bilansowania, lasy losowe, wynik modelu