IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Machine learning-based approaches for automatic data validation and outlier control of loan microdata in the Bank of Russia[1]

Dmitrii Diachkov,
Bank of Russia

[1] This presentation was prepared for the Workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the central banks and other institutions represented at the event.

Machine learning-based approaches for automatic data validation and outlier control of loan microdata in the Bank of Russia

Dmitrii Diachkov[1]

Abstract

Validation of loan microdata is an outstanding issue in central bank statistics, and the challenge is magnified by increasing variability and heterogeneity of the underlying data. In this work, an application of machine learning approaches in large loan datasets is discussed. We identified a set of tools, such as gradient boosting, neural networks, random forests, that can be used to enhance the quality of microdata on loans available in the Bank of Russia. Ensemble methods and pre-processing techniques in RStudio are used to explore, analyse and determine outliers and potentially erroneous clusters of data on loans and borrowers. Based on the ensemble of machine learning algorithms, the toolkit efficiently reduces the variance of predictions and, in some cases, outperforming base classifiers (logistic regression) is expected to be very useful for quality control. The results reveal that highly atypical groups may be identified, providing additional insight for further scrutiny, methodological research, and the development of statistical indicators.

Keywords: machine learning, data validation, outliers, bank loans, data quality

JEL classification: E51, E58, G18, G21, C81

[1] Dmitrii Diachkov is an Expert Team Leader in Statistics Department at the Bank of Russia and a Statistical Systems student of NOVA IMS, Lisbon, Portugal. I would like to thank my colleagues from the Bank of Russia's Statistics Department for their valuable ideas, support and assistance. The views expressed are those of the author and not necessarily those of the Bank of Russia or NOVA IMS.

Contents

Introduction

Loan-level information is one of the primary data sources for a wide variety of analytical tasks performed in central banks (Morandi, Nicoletti, 2017). The use of highly granular data (microdata) opens up new frontiers for analyzing the dynamics and structure of the credit market (Santos, 2013).

Microdata on loans can be used for various purposes like a compilation of monetary statistics (Dyachkov, Nurimanova, 2017), sectoral credit risks analysis, or supporting decisions on the countercyclical capital buffer (Carstens, 2016) in order to analyze actual distributions of indicators, and not their distorted representations by aggregation (Aaron & Hogg, 2005).

Despite all the advantages of microdata, some factors hinder their full use (Livraga, 2019), including possible quality problems (Osiewicz, Fache-Rousova & Kulmala, 2015). Even though the data obtained from administrative sources are, in fact, of relatively higher quality (Crato & Paruolo, 2018), in order to maximize their usefulness, verification procedures of the incoming data should be built. The use of machine learning will reduce the cost of producing statistics and improve its quality (Tam & Clarke, 2015).

The paper analyzes existing approaches to classification in large volumes of microdata using machine learning methods, proposes new ways to solve applied problems for analyzing large volumes of data to improve their quality and search for atypical values. The main advantages and disadvantages of using decision trees, regression trees, and random forests in classification problems are considered, several models for the practical application of machine learning methods are proposed, including gradient boosting methods. Particular attention is paid to the problems of balancing the training samples of microdata.

The microdata sets on loans available to the Bank of Russia contain more than 150 attributes and cover 100% of all loans to legal entities and individual entrepreneurs provided by banks in Russia.

Due to limited resource capacity and rapidly increasing data volumes, analysis of the reliability of a large and diverse set of data can become a tricky task; hence it simultaneously allows the application of machine learning methods to amplify the efficiency of data quality control and decisions made on their basis.

Interpretability, controllability, the possibility of automatic selection of informative features of decision trees, and regressions were the reason for their use as the primary tool for efficient classification of processing large amounts of data, searching for atypical values for subsequent filtering and identifying erroneous values in categorical variables. Current research is made to improve data quality and is based on a variety of disciplines, and represents a rich set of scientific and technological tasks for statisticians.

Literature review

Data availability alone does not guarantee that assumptions, derived from this data, are correct as well as it does not guarantee that data management functions are efficient (Manjunath, Hegadi, Ravikumar 2010).

The issue is not new, but it would be safe to estimate that a large-scale numeric database without critical judgment can have an error rate of 2-5% (Dong et al., 2002); hence it is not clear what can be considered as acceptable accuracy. According to Karr et al. (2006), some time ago, data quality was just a scientific problem rooted in measurement errors and research uncertainty.

Nevertheless, in today's world of high dimensional data and complex economic policy decisions, data quality problems can create significant economic losses (Madhikermi et al., 2016) and short-term fixes (Lee et al., 2006), and organizations tend to underestimate the consequences that in fact may vary "from significant to catastrophic" (Gudivada, Apon & Ding, 2015). Errors in microdata increase variance and create biased results (Eurostat, 2020).

Consumers can evaluate data quality based on their objectives (Lee et al., 2006). The same datasets may be used for multiple tasks that need different quality characteristics (Batini et al., 2015 and Aljumaili, 2016). In addition, if task requirements change over time, some quality characteristics might change (Eurostat, 2007). Variables in large microdata databases that are not of particular interest for data owners may be of lower quality (Crato & Paruolo, 2019). Therefore, providing high-quality data means tracking a constantly moving target. Perrella & Catz (2020) conclude that IT tools should not only provide consistency checks but assess the plausibility of the data.

Gomolka et al. (2021) conclude that it is essential to track the results of data quality checks to make sure that any modifications conducted to enhance data quality for researchers do not affect the contents of data. Maintaining the high quality of a microdata register might become a challenging task because of small but frequent objects, such as taxpayers' data (Gavin, 2021).

All these challenges can be resolved with appropriate work-process organization and modern computer science methods (Crato & Paruolo, 2019), taking into account a context-based nature of data quality (Batini et al., 2015).

The main goal of modern microdata quality control is protection from incorrect or invalid information (Crato & Paruolo, 2019) and missing data (Smith et al., 2018). Eliminating and rectifying these quality gaps should be one of the primary concerns for statisticians (Perrella & Catz, 2020), and internal researchers need microdata to be as precise as possible to achieve reliable results (Domingo-Ferrer & Blanco-Justicia, 2021).

We live in the "era of big data" and collecting such data requires tremendous effort, and publication is often delayed. However, there has been an explosive growth in the amount of data available to use (Doerr, Gambacorta & Garralda 2021). New data collection and dissemination models enable real-time analysis of massive amounts of data.

The growing use of artificial intelligence (AI) and machine learning applications makes it even more difficult to ensure data quality in organizations (Janssen et al., 2020) as well as the introduction of real-time streaming platforms that continuously transmit large amounts of data to corporate systems. In addition, data quality now often needs to be managed in combination with on-premises and cloud systems, and hence data quality enhancement tools should be embedded in the process of quality management (Kropf, 2020).

Usually, microdata quality control is designed by implementing automatic checking procedures (logical or mathematical rules) during the data collection (Zambuto et al., 2020).

The literature on data quality has not yet paid proper regard to the design of methodologies that can provide automatized verification of large multivariate datasets (Farnè & Vouldis, 2018). However, effective ML applications require high-quality datasets. For example, data that is distorted by outliers can result in non-convergence in ensemble learning and a dramatic reduction of quality in prediction (Gudivada, Apon & Ding, 2015).

Coeuré (2017) states that carrying out automatic checks, based on machine learning techniques, and AI is one of the ways to ensure that data remain of high quality. According to Zambuto et al. (2020), machine-learning application to microdata will be more efficient when designed models are backed up by pre-determined relationships, such as accounting rules. The predictive power of all ML models improves when the length of the relationship between attributes and objects increases, and the non-linear relationship between variables in a dataset is better mined my machine learning (Gambacorta et al., 2019).

Using data on proprietary loan transactions from a leading fintech company in China, Gambacorta et al. (2019) show that while regular models perform well in ordinary times, machine learning models are way better able to predict extraordinary events. It is interesting to note that a possible reason for that behavior is that machine learning can better react to the non-linear relationship between factors. Lukauskas & Ruzgas (2021) used various techniques like artificial neural networks, XGBoost, LightGBM, Catboost to predict borrowers' default taking into account computational time. The recent research by Severino & Peng (2021) revealed that ensemble-based random forests, gradient boosting, and neural networks achieve the most effective results, overcoming base classifiers such as logistic regression. Likewise, Dou et al. (2019) analyzed online fraud and applied the XGBoost model to predict fraud using a variety of feature sets, achieving more than 99% level of accuracy, taking into account the prediction class balance. Odegua (2020) has successfully used the XGBoost for bank loan default prediction.

Regular decision trees are intuitive because the model visualizes a decision scheme (Wang et al., 2020). Many researchers support the idea because of the high interpretability and flexibility in feature selection of decision trees (Lee et al., 2006; Kao et al., 2012).

Divakar & Chitharanjan (2019) explored credit card fraud data using several boosting methods and concluded that the XGBoost algorithm is the best model among others considered, like AdaBoost and Gradient Boost. Raju (2021) shows that the method may outperform random forests and OLightGBM models, while Trisanto et al. (2021) showed that imbalanced data might be considered an issue to this method.

Manjeet et al. (2018) compared traditional classification models with neural networks concluded that the neural network other models to predict loan default. Li Ying (2018) compared three model families: random forest, logistic regression, and SVM on bank loan data. The results show that random forests generally perform better. This result corresponds with a study of COŞER, Maer-matei & ALBU (2019) that compared LightGBM, XGBoost, Logistic Regression, random forest to evaluate loan probability default and recognized random forest as an optimal classifier for the task. The learning rate is fast and can be applied to large-scale datasets (Wang et al., 2020)

According to Cheng (2021), XGBoost training requires cumbersome setting and adjustment; hence Koduru et al. (2020) offer even more complicated ML tools such as random forest + XGBoost for loan application scoring.

Motivated by all presented research, we explore microdata on loans in the Bank of Russia and seek possible application of ML methods to enhance data quality of the essential attributes.

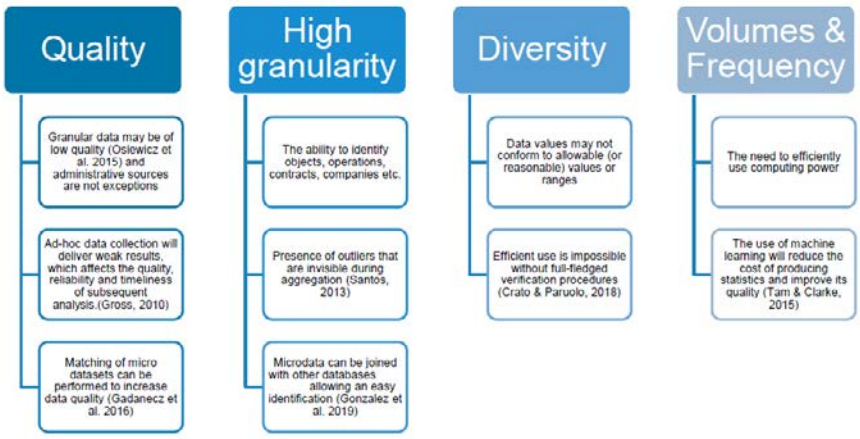
Loan microdata quality control in the Bank of Russia

Magnified by variability and heterogeneity of the underlying data, validation of loan microdata is an outstanding issue in the Bank of Russia.

Practical difficulties in collecting and processing microdata, non-transparent methodology, lack of quality assessments, and instability of "big data" do not allow putting these new sources of information on a par with data used in normal statistical business processes. This situation requires conceptual comprehension and evolutionary development with the approbation of approaches on individual projects.

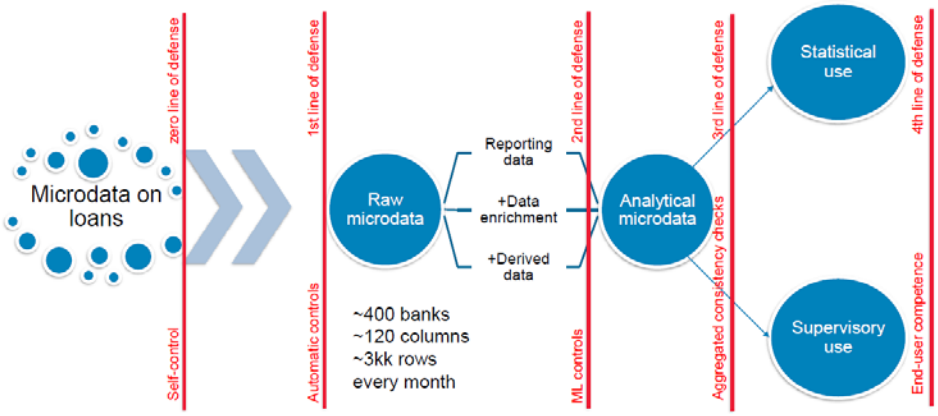
The accumulated experience of working with loan microdata in the Bank of Russia's Statistics Department is devised based on four main data quality components presented in Figure 1.

Figure 1: Accumulated experience of working with loan microdata



The conceptional scheme of loan microdata pipeline in the Bank of Russia is presented in Figure 2. The best way to apply ML for data quality is before or when forming analytical microdata. All data items should be enriched at this stage, and derived columns calculated and applying ML controls may potentially form the 2nd line of defence (after automatic controls, but before aggregated consistency checks).

Figure 2: Possible implementation of ML controls to data quality control scheme

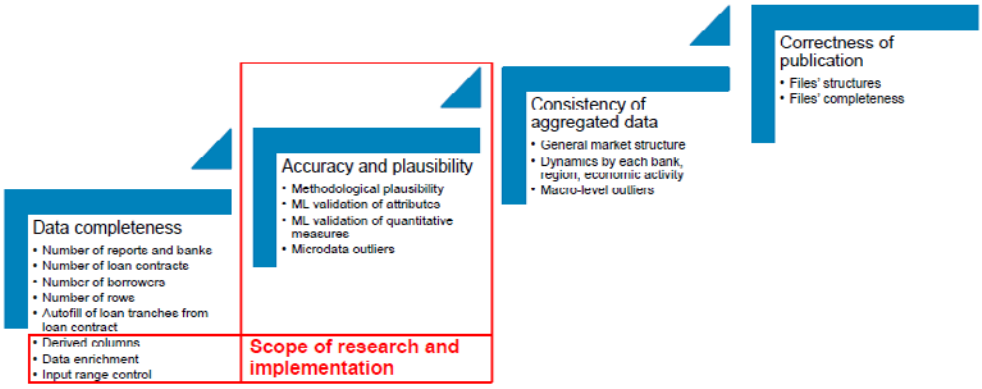


Examples of errors that arise in the process of collecting information are the following:

- mistyping errors;
- technical errors (i.e., numerical field that contains 0s, that can be economically interpreted, for missing values);
- errors of database merges (i.e., a big bank with comprehensive branch coverage collects data separately from each branch and unite that data in its' warehouse)
- unit conversion errors (i.e., thousand of units versus units);
- errors of source data interpretation (i.e., misreading of the original documents);
- errors in the compilation of the metadata (i.e., the supporting information is incorrectly entered or interpreted);
- errors in the underlying data (i.e., the results presented in the original data source are incorrect or misleading).

All the error types mentioned above may be found at any stage of data processing. The possible niche for implementing first ML models and their role in verification workflow is presented in Figure 3 below.

Figure 3: Scope of first possible ML controls in a context of data verification



The main goal of adding the ML component to this workflow is to explore the possibility of applying relatively simple machine learning methods to improve the overall quality of microdata and decisions based on them in the Bank of Russia.

Limited human abilities (analysts or statisticians) to extensively analyze the reliability of large and complicated datasets, furthermore changing over time, inevitably leads to ML applications. The effective application should meet the following criteria:

Criteria for effective ML application

Table 1

Criterion	Purpose
High interpretability of results	All participants of data quality management and end-users should be able to understand the principles of control and resulting outputs.
Moderate ease of implementation	Applications should be relatively simple so staff members without strong IT knowledge could implement them in the field of competence
Moderate process control	Underlying banking data is constantly changing so applied techniques should be amendable and provide relatively robust results when methodology of data collection changes or new products emerge
High scalability and reusability	Good models or model patterns should be re-used (if appropriate) for similar data types as well as be independent of training dataset size.
High automation capability	Quality control needs to be designed that way so it can be executed on server instance by timetable or by request.

Empirical applications of ML quality controls

In order to create a new line of data quality validation, the following datasets were used as material for approbation: Banks' reporting form 0409303 "Information on loans granted to legal entities and individual entrepreneurs" with microdata on loans, annual accounting data, Statistical Registry on companies provided by State Statistics Service, State Registry of SMEs provided by Federal Tax Service.

For proper validation of underlying data for loan statistics in the Bank of Russia, it is vital to pay close attention to the industrial classification of borrowers as one of the main focal points as well as SME classification and main loan parameters. The design of the first models should be targeted at main grouping variables or attributes used to form input data arrays.

1. Validation of balance sheet codes for borrowers with decision trees, XGBoost, and logistic regression

Balance sheet code is a simple and understandable attribute for economists that can be used to filter borrowers by type and industry. Information about the balance sheet codes is reported to the Bank of Russia by banks directly. Balance sheet codes may be determined by banks based on irrelevant data or just by mistake. Another type of mistake is technical errors during report preparation or submission.

The bank should assign the balance sheet code to its borrower based on a limited list of parameters, such as type of loan, business entity forms, and economic activity, so according to the factors mentioned earlier, we can verify balance sheet codes with the data of public registers or other data sources. Generally, we would like to have a

probability score of correct balance sheet code for each item as an output from our ML model.

Building on Wang et al. (2020), Lee et al. (2006), Kao et al. (2012), we developed a set of regular decision trees to promote interpretable and easily visualized models for binary classification. The rationale behind these models was based on the idea that a combination of loan type, economic activity, business entity form, and debt sum may be used to answer the question of whether the reported balance sheet code is valid or not.

Standard balance sheet codes that are used for loan statistics and hence to be validated in a dataset with 1056k observations:

«452» - Loans to non-financial companies;

«454» - Loans to individual entrepreneurs;

«451» - Loans to financial companies.

Both «452» and «454» sub-samples did not require additional transmutations, and the dependent variable was evenly distributed in the training and test samples. We developed a set of 15 models (solutions for account «452» presented in Figure 4 in Appendix) with different specifications to seek the best dependence between response and explanatory variables.

Since the training and test samples are balanced (predictable classes are evenly distributed), standard measure like accuracy (1) is an excellent metric of model quality.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

As many models have led to relatively comparable results with high accuracy, we decided to introduce computational time as a penalty for efficiency, normalized by the number of explanatory variables (Figure 5a in Appendix). Hence, relative cost-efficiency for the evaluation purpose of several models with balanced classes could be formulated as (2):

$$\boldsymbol{Relative\ Efficiency} = \frac{\boldsymbol{E(Accuracy)}}{\boldsymbol{E(Time)}} \quad (2),$$

We also tested XGBoost (Figure 5b in Appendix) and logistic regression models to predict the same balanced classes, and it achieved the same results in terms of efficiency but with a longer computational time. Results are presented in Table 2.

Summary of results for decision tree, XGBoost, and logistic regression approaches to classification task for account code "452"

Table 2

Model type	Mean accuracy	Mean precision	Mean recall	Mean time of fit and predict (per 1kk rows), seconds	Relative efficiency score
Decision tree	0,989	0,987	0,990	17	1,939
XGBoost	0,988	0,993	0,984	45	0,731
Logistic regression	0,989	0,99	0,989	113	0,291

During the analysis of the model output, we did a deviation analysis. It showed that the decision tree classified 0.66% or 1331 loans as loans of non-financial companies (code «452»), but in fact, they are reported as other codes (FP, for example, «453» loans to non-residents or «451» loans to financial companies). Additionally, 0.44% or 983 loans were classified by the decision tree as loans that should be recorded as non «452», but in fact, they are recorded in «452» accounts (FN, mainly financial companies, which should be reflected with code «451»). These deviations should be considered as rare events or errors and subsequently scrutinized. The reason behind this atypical coding may be one of the following factors:

- Methodological features, which should lead to the methodology refinement;
- Reporting errors, which leads to informing the bank about the error and asking for report re-submission;
- Accounting errors, which require to inform the bank about the error and elimination in the future.

Luckily, we have not faced significant challenges handling imbalanced classes. While codes «452», «454» represent almost a half of all observations each, «451», corresponding to financial companies, was presented only in 1-2% of observations. The dependent variable is not evenly distributed in the training and test samples.

A similar set of 15 models were derived, and the same computational limitations were implemented as well as in the case of balanced samples, but they were adapted for imbalanced samples and normalized by the number of explanatory variables. Hence, relative cost-efficiency for the evaluation purpose of several models with imbalanced classes could be formulated as (3):

$$\boldsymbol{Relative\ Efficiency2} = \frac{E(F1\_score)}{E(Time)}\ (3),$$

Precision (4) and Recall (5) metrics do not depend on the ratio of classes and therefore are applicable in conditions of imbalanced samples. The F1 score is a good balance between these two metrics.

$$\boldsymbol{Precision} = \frac{TP}{TP \mid FP}\ (4) \quad \boldsymbol{Recall} = \frac{TP}{TP+FN}\ (5)$$

$$\boldsymbol{F1\_score} = \frac{2 * Precision * Recall}{Precision + Recall}\ (6)$$

The affiliation of the legal entity to the finance industry «K» and specific loan types both determine the belonging to the code «451». Despite the high accuracy (99.1%), the confidence in the model was significantly lower since the training and test samples were unbalanced. Accuracy in this situation is an incorrect metric. However, the disproportion of classes in the task does not influence the overall performance presented in Table 3 below.

Summary of results for decision tree, XGBoost, and logistic regression approaches to classification task for account code "451"

Table 3

Train data type	Model type	Accuracy, %	Fit and predict time (seconds per 1kk rows)	F1_score, %	Relative efficiency score 2
Downsampled (30k rows)	Decision tree	98,98%	9,976	99,48%	0,100
	Logistic regression	99,35%	62,34	99,67%	0,016
	XGBoost	99,35%	50,606	99,67%	0,020
Original - imbalanced (845k rows)	Decision tree	99,43%	14,081	99,71%	0,071
	Logistic regression	99,44%	136,62	99,71%	0,007
	XGBoost	99,32%	38,783	99,65%	0,026
Upsampled (1658k rows)	Decision tree	98,99%	15,331	99,48%	0,065
	Logistic regression	99,39%	125,16	99,69%	0,008
	XGBoost	99,35%	70,38	99,67%	0,014

We can conclude that in this particular case, class imbalance does not affect model performance. In such cases, we should shift towards down-sampled versions because of fewer computational expenditures.

2. Validation of interest rates data with XGBoost

Interest rate statistics require only high-quality data because a single outlier can dramatically change weighted average rates. To verify outliers, we decided to develop an algorithm to define a pattern of ordinary data items. XGBoost provided a quick and effective solution.

Using data on 421k loans as a training dataset and 105k loans as a test dataset, the XGBoost algorithm was implemented and benchmarked with neural net and regular linear regression (Table 4).

Comparison results

Table 4

Model	RMSE
XGBoost	2,04
Neural net (caret and nnet)	3,25
Linear Regression	3,85

Nine variables were used as an input (Figure 6 in Appendix), but subsequently, the number was reduced due to little importance of some variables.

XGBoost model showed better results, dynamics of RMSE of each training repetition, and testing presented in Figure 7 (Appendix). It can be used to restore omissions in interest rates and search for outliers. The corresponding actual VS XGBoost predicted values for interest rates by loan type (as a top-importance variable) are presented in Figure 8 in Appendix.

Another critical task for interest rate statistics is a breakdown of loans by the size of the borrower's business. Hence, XGBoost multiclass classifier was used to predict the size of SME companies: micro, small, medium, or non-SME.

Two approaches were used: XGBoost on balanced train data and imbalanced. The results were very similar, but the balanced sample provided more insight into the importance of features. However, if the importance of features does not matter, one can apply the technique without up-sampling (to save computational resources) or down-sampling (because it lessens the model's exposure to non-standard objects). Confusion matrices and feature importance structures are presented in Figure 9 in Appendix.

Finally, we conclude that XGBoost is a powerful tool with high speed and very high accuracy on millions of rows. Performance on the imbalanced sample is 2% less accurate than on the upsampled and balanced (90% VS 92%).

3. Validation of SME status with neural networks, random forests and logistic regression

Belonging to the SME Registry (provided by Federal Tax Service) defines a borrower's business size. This attribute is crucial for the analysis of the economic situation. However, small and insignificant companies may be excluded from the Registry for various reasons. We need to be able to check the validity of any given status.

Large borrowers have more assets and, accordingly, apply for more significant amounts of loans. SME borrowers are usually smaller in business size and balance sheet.

To solve this task, we have built several simple neural networks that classify companies as SMEs and non-SMEs based on various quantitative indicators characterizing the loan and the borrower. The dependent variable is not evenly distributed in the training and test samples (75% of borrowers are SMEs), but this issue was solved with down-sampling. Down-sampling was the only option because of already vast amounts of data items. We compared 20 different compositions of neural networks. The best neural network consisted of 2 hidden layers (with 4 and 3 neurons respectively – figure 11) and predicted SME status with 90,4% accuracy and F1_score of 80% (figure 12). This model cannot be considered ready for productional usage in data quality control because of the time-consuming training process and lack of precision and recall, which results in a low F1_score with many false-positive SME-statuses, so the model needs to be re-designed.

We approached the same task with randomForest and logistic regression and achieved the same accuracy and F1_score results (randomForest confusion matrix is presented in Figure 13 of the Appendix), but faster. Additional implementations were made with six explanatory variables on random sampling from up-sampled (balanced) training data. In terms of computational speed, the efficiency of random forest is two times higher, while results are even slightly better. Results are presented in Table 5 below. Traditional classifier logistic regression has beaten neural networks and random forests in terms of result/speed ratio.

Comparison results Table 5

Model	Accuracy	F1_score	Computation time (per 1kk rows), seconds
Neural net	90,4%	80,1%	757
Random forest	92,8%	81,4%	318
Log regression	93,3%	83,1%	108

Conclusions

The main lessons are following:

Interpretability, controllability, the possibility of automatic selection of informative features of decision trees, and regressions were the reason for their use as the primary tool for efficient classification of processing large amounts of data, searching for atypical values for subsequent filtering, and identifying erroneous values in categorical variables.

Due to human disabilities, to analyze the reliability of a large and diverse set of data, expanding the field of applied machine learning methods will increase the quality of data and decisions made on their basis.

When solving classification problems, metrics should be monitored carefully and problems solved under business logic. With unequal classes, metrics should be selected carefully, and up-sampling or down-sampling applied when necessary.

Simpler models often give more balanced and correct results during cross-validation on test data.

Appendix

Figure 4. Efficiency of 15 different specifications for decision trees, predicting the class «452» (non-financial companies) after 100 repetitions (green highlight – best models).

Formula	Mean accuracy, %	Mean time (seconds per 1kk rows)	Relative Efficiency
~ ECON_ACTIVITY + BORROWER_TYPE + LOAN_TYPE	98,904	17,044	5,80
~ ECON_ACTIVITY + LOAN_TYPE + DEBT + BORROWER_TYPE	98,886	18,257	5,42
~ BORROWER_TYPE + LOAN_TYPE	97,352	19,742	4,93
~ BORROWER_TYPE + LOAN_TYPE + DEBT	97,346	20,512	4,75
~ ECON_ACTIVITY + BORROWER_TYPE + DEBT	96,064	22,028	4,36
~ ECON_ACTIVITY + BORROWER_TYPE	95,981	19,546	4,91
~ BORROWER_TYPE	94,182	17,071	5,52
~ BORROWER_TYPE + DEBT	94,155	23,28	4,04
~ ECON_ACTIVITY + LOAN_TYPE	69,813	22,616	3,09
~ ECON_ACTIVITY + LOAN_TYPE + DEBT	69,732	35,284	1,98
~ LOAN_TYPE	67,244	16,122	4,17
~ LOAN_TYPE + DEBT	67,127	30,853	2,18
~ ECON_ACTIVITY	61,317	17,175	3,57
~ ECON_ACTIVITY + DEBT	61,226	32,287	1,90
~ DEBT	54,282	28,043	1,94

Figure 5a. Best decision tree structures and confusion matrices for predictions of class «452» (non-financial companies)

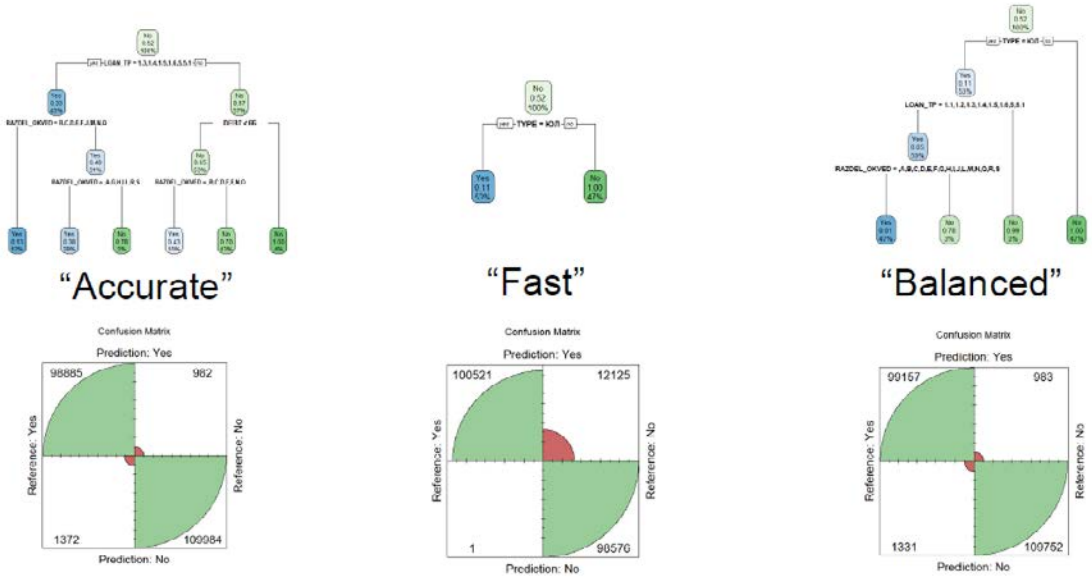


Figure 5b. Confusion matrix (left) and feature importance for XGBoost model for predictions of class «452» (non-financial companies)

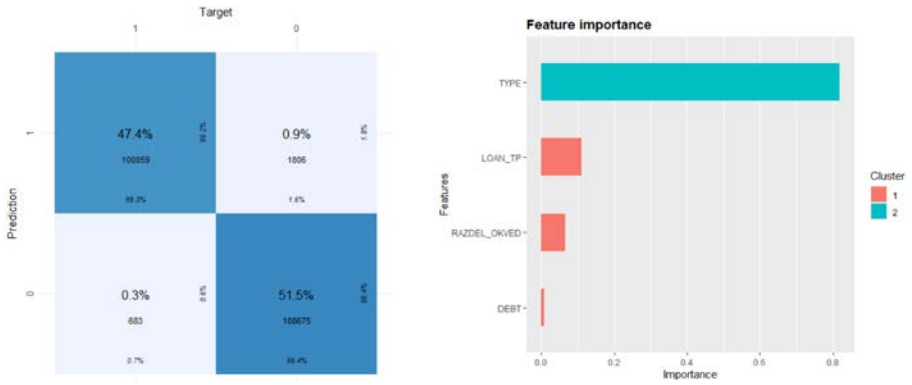


Figure 6. Importance matrix of XGboost models for interest rate prediction

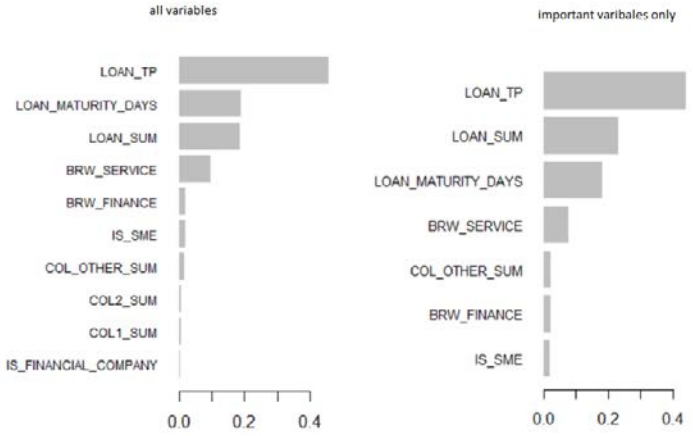


Figure 7. RMSE for XGboost interest rate prediction (train and test)

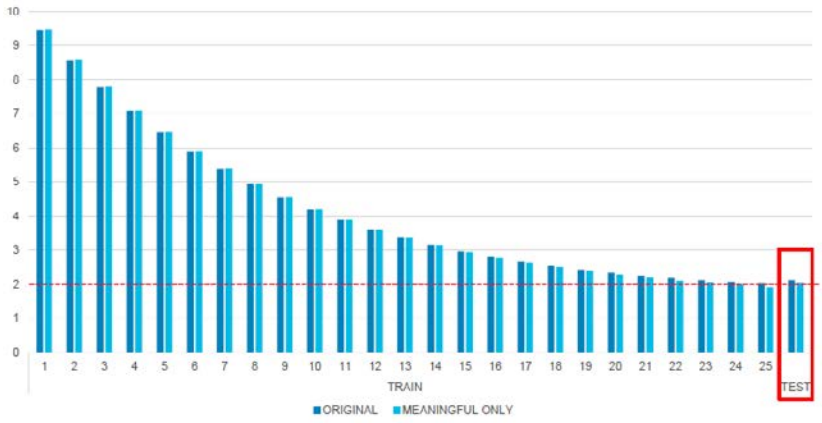


Figure 8. Correspondence of XGBoost predicted (y_pred) and actual (y) values for interest rates by loan type (as top-importance variable)

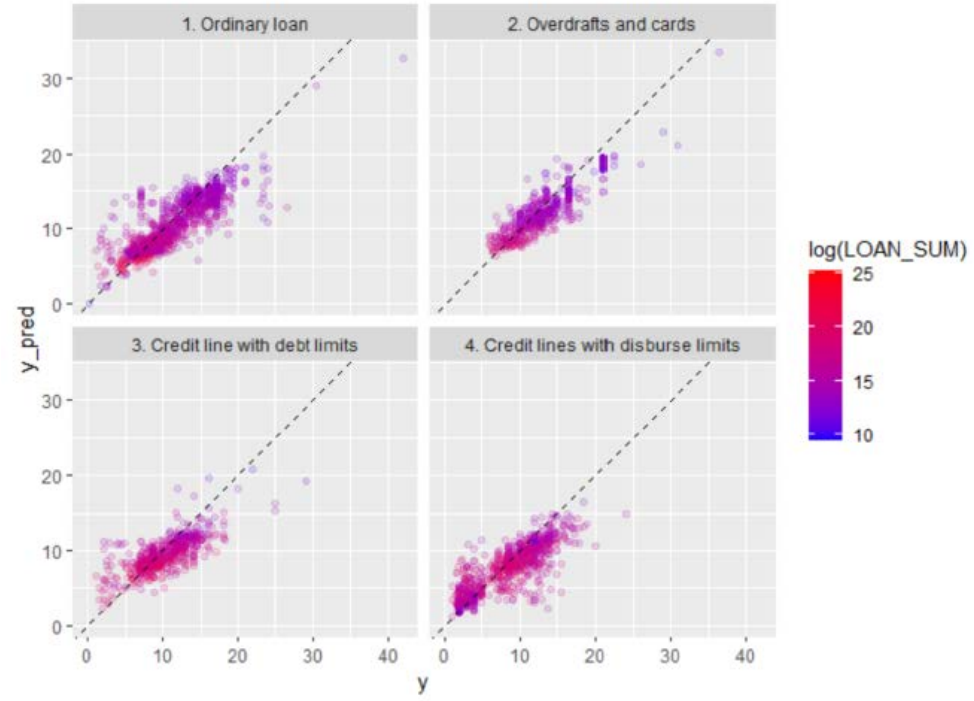


Figure 9. Confusion matrices for XGBoost predictions of a company's size (where 1 – is micro, 2 – small, 3 – medium, 4 – non-SME)

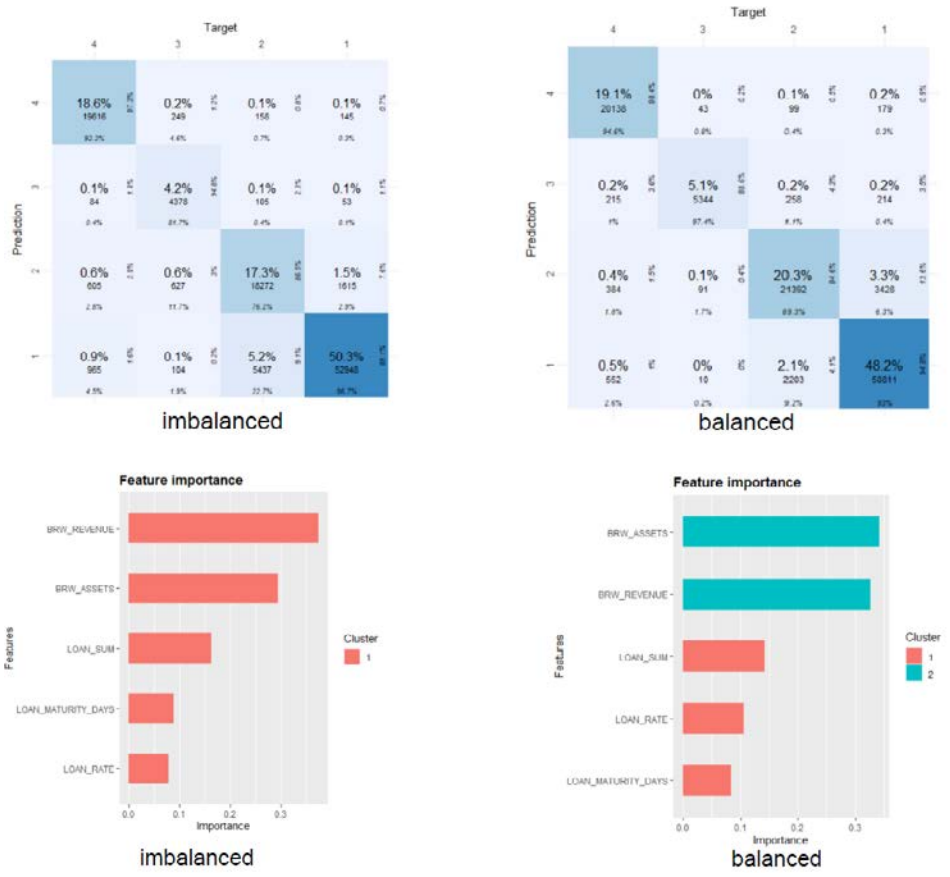


Figure 10. Accuracy and fitting time of tested neural networks with different composition of layer and neurons to predict SME status (green highlight – the best model).

Hidden layers	Accuracy	Fitting time	Prediction time	Weighted score
4,3	90,4%	619,78	138,86	10,77211413
2,2	67,2%	102,54	334,39	10,33535241
3,2	73,2%	232,73	299,22	10,07282431
4	79,8%	518,30	245,60	8,336222782
5,3	71,0%	339,18	305,19	7,823119106
4,2	71,2%	264,82	420,88	7,393118623
4,1	80,6%	607,07	346,24	6,814520418
5,3	58,2%	7,68	493,28	6,761507372
3	78,8%	484,72	444,25	6,684266136
4,1	47,4%	5,19	337,11	6,563785262
4,3	55,8%	11,47	539,54	5,650794018
4,2	70,8%	619,94	354,98	5,141586833
2	77,6%	1063,63	156,36	4,935904051
3,2	88,4%	1398,23	281,46	4,652387824
5	69,4%	321,55	760,02	4,453117059
4,3	43,8%	6,85	426,56	4,426365879
4,2	70,6%	796,28	380,96	4,233930891
5,4	76,8%	1431,68	377,27	3,260586787
6,2	77,8%	1619,66	292,87	3,16482278

Figure 11. Best neural network for SME size prediction

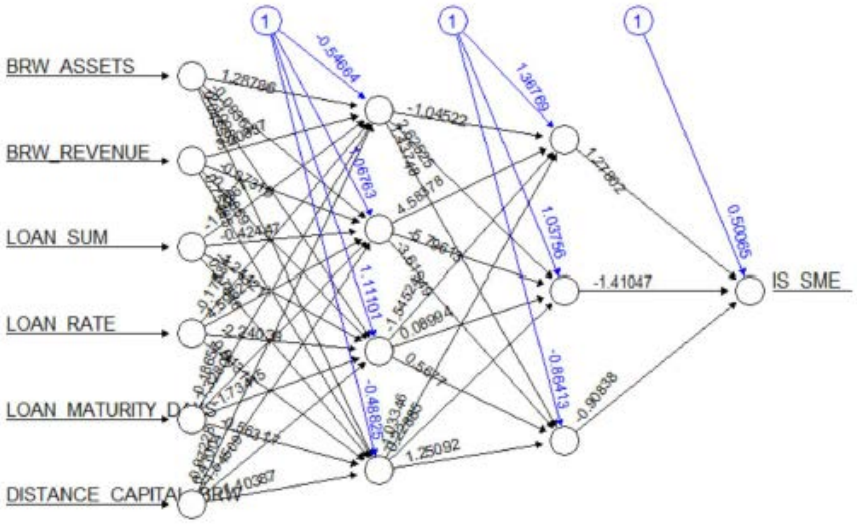


Figure 12. Confusion matrix for the neural network for SME-status prediction

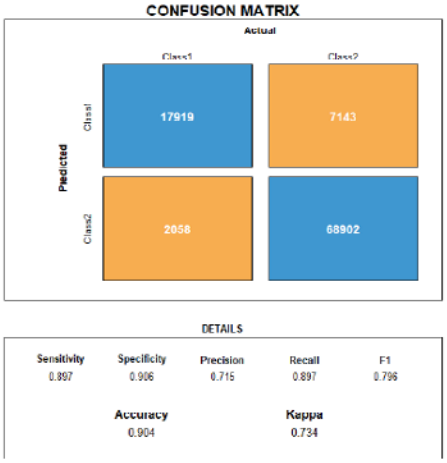
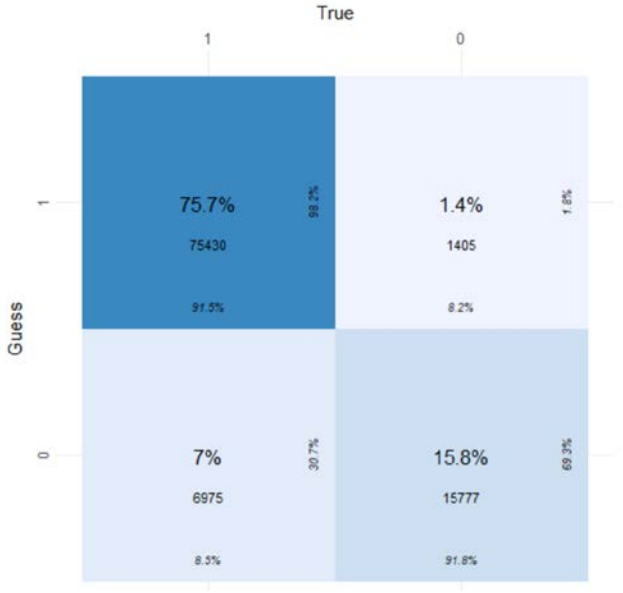


Figure 13. Confusion matrix for random forest solution for SME prediction



References

1. Aaron M., Hogg D. (2005) The use of microdata to assess risks in the non-financial corporate sector, Financial System Review, December, Bank of Canada.

2. Aljumaili, M. (2016). Data quality assessment: Applied in maintenance (Doctoral dissertation, Luleå tekniska universitet).

3. Batini, Carlo & Rula, Anisa & Scannapieco, Monica & Viscusi, Gianluigi. (2015). From Data Quality to Big Data Quality. Journal of Database Management. 26. 60-82. 10.4018/JDM.2015010103.

4. Blaise Gadanecz & Bruno Tissot & Mariagnese Branchi & Mario Ascolese, 2016. "The sharing of micro data – a central bank perspective," IFC Reports 6, Bank for International Settlements.

5. Carstens A. (2016) Micro-data as a Key Input to Designing Macro-prudential Policy: The Mexican Experience // Eighth European Central Bank Conference on Statistics, p.1-18

6. Carstens, A. (2016). Micro-data as a Key Input to Designing Macro-prudential Policy: The Mexican Experience. Remarks at the Eighth European Central Bank Conference on Statistics

7. Cheng, Y. (2021, April). Research on Credit Strategy Based on XGBoost Algorithm and Optimization Problem. In Journal of Physics: Conference Series (Vol. 1865, No. 4, p. 042137). IOP Publishing.

8. Coeuré, B., "Setting standards for granular data", opening remarks at the Third OFR-ECB-Bank of England workshop on "Setting Global Standards for Granular Data: Sharing the Challenge", Frankfurt am Main, March 2017.

9. COŞER, A., Maer-matei, M. M., & ALBU, C. (2019). PREDICTIVE MODELS FOR LOAN DEFAULT RISK ASSESSMENT. Economic Computation & Economic Cybernetics Studies & Research, 53(2).

10. Crato, N. & Paruolo, P. (Eds.). (2018). Data-driven policy impact evaluation: How Access to microdata is transforming policy design. Springer.

11. Crato, N., & Paruolo, P. (2019). The Power of Microdata: An Introduction. In Data-Driven Policy Impact Evaluation (pp. 1-14). Springer, Cham.

12. Divakar, K., & Chitharanjan, K. (2019). Performance evaluation of credit card fraud transactions using boosting algorithms. Int. J. Electron. Commun. Comput. Eng. IJECCE, 10(6), 262-270.

13. Domingo-Ferrer, J., & Blanco-Justicia, A. (2021, September). Towards Machine Learning-Assisted Output Checking for Statistical Disclosure Control. In International Conference on Modeling Decisions for Artificial Intelligence (pp. 335-345). Springer, Cham.

14. Dong, Q., Yan, X., Wilhoit, R. C., Hong, X., Chirico, R. D., Diky, V. V., & Frenkel, M. (2002). Data Quality Assurance for Thermophysical Property Databases Applications to the TRC SOURCE Data System. Journal of chemical information and computer sciences, 42(3), 473-480.

15. Dou, Y., Li, W., Liu, Z., Dong, Z., Luo, J., & Philip, S. Y. (2019, August). Uncovering download fraud activities in mobile app markets. In 2019 IEEE/ACM International

Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 671-678). IEEE.

16. Dyachkov, D.V, Nurimanova, I.F. (2017) Specifics of microdata-based statistics of interest rates on lending to non-financial sector // Russian Journal of Money and Finance (Money and Credit). 2017 Issue 12 – p. 64-72.

17. Eurostat (2007), Handbook on Data Quality Assessment Methods and Tools. Editors: Manfred Ehling and Thomas Körner.

18. Eurostat (2020). European Statistical System (ESS) handbook for quality and metadata reports — 2020 edition.

19. Farnè, Matteo; Vouldis, Angelos T. (2018) : A methodology for automatised outlier detection in high-dimensional datasets: An application to euro area banks' supervisory data, ECB Working Paper, No. 2171, ISBN 978-92-899-3276-9, European Central Bank (ECB), Frankfurt a. M.

20. Gambacorta, L., Huang, Y., Qiu, H., & Wang, J. (2019). How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm.

21. Gavin, E. (2021). How to Collaborate Effectively to Improve Data Quality and Use in Revenue Administration and Official Statistics. IMF How To Notes, 2021(005).

22. Gomolka, M., Blaschke, J., Brîncoveanu, C., Hirsch, C., & Yalcin, E. Data Orchestration Blueprint Based on YAML {dobby} Research data pipelines in R.

23. González A.G. & Valadez M.S. & Cerecero M.R, 2019. "Sharing and using financial micro-data," IFC Bulletins chapters, in: Bank for International Settlements (ed.), Are post-crisis statistical initiatives completed?, volume 49, Bank for International Settlements.

24. Gross, F. (2010). Micro-data as a necessary infrastructure–standardisation of reference data on instruments and entities as a starting point: need for a Reference Data Utility. IFC Bulletin, 25, 334.

25. Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. Government Information Quarterly, 37(3), 101493.

26. Kao, L.-J., Chiu, C.-C., and Chiu, F.-Y. (2012). A bayesian latent variable model with classification and regression tree approach for behavior and credit scoring. Knowledge-Based Systems, 36:245–252.

27. Karr, Alan F., Ashish P. Sanil, and David L. Banks. "Data quality: A statistical perspective." Statistical Methodology 3.2 (2006): 137-173.

28. Koduru, M., Pranati C., M., Phanidhar, M., & Srinivas, D. K. (2020). RF-XGBoost Model for Loan Application Scoring in Non Banking Financial Institutions. International Journal of Engineering Research & Technology (IJERT) ISSN, 2278-0181.

29. Kropf, S. L. (2020, November). ENHANCING DATA QUALITY FOR DATA ANALYTICS THROUGH MACHINE LEARNING. In European Scientific Conference of Doctoral Students (p. 97).

30. Lee, Y. W., Pipino, L., Funk, J. D., & Wang, R. Y. (2006). Journey to data quality (pp. 137-150). Cambridge: MIT press.

31. Li Ying. (2018). Research on bank credit default prediction based on data mining algorithm. The International Journal of Social Science and Humanities Invention 5(06): 4820-4820, ISSN: 2349-2031.

32. Livraga, G. (2019). Privacy in microdata release: Challenges, techniques, and approaches. In Data-Driven Policy Impact Evaluation (pp. 67-83). Springer, Cham.

33. Lukauskas, M., & Ruzgas, T. (2021). Bank credit card default classification based on clustering using machine learning algorithms. In 9th world sustainability forum, virtual, Switzerland, 13–15 September 2021: program and abstract book. MDPI.

34. Madhikermi, M., Kubler, S., Robert, J., Buda, A., & Främling, K. (2016). Data quality assessment of maintenance reporting procedures. Expert Systems with Applications, 63, 145-164.

35. Manjeet K., Vishesh G., Tarun J., Sahil S., DR. Lalit M. G. (2018). Neural Network Approach To Loan Default Prediction, International Research Journal of Engineering and Technology (IRJET) , p-ISSN: 2395-0072

36. Manjunath, T. N., Hegadi, R. S., & Ravikumar, G. K. (2010). Analysis of data quality aspects in datawarehouse systems. International Journal of Computer Science and Information Technologies, 2(1), 477-485.

37. Morandi, G., & Nicoletti, G. (2017). Using microdata from monetary statistics to understand intra-group transactions and their implication in financial stability issues. IFC Bulletins chapters, 46.

38. Odegua, R. (2020). Predicting Bank Loan Default with Extreme Gradient Boosting. arXiv preprint arXiv:2002.02011.

39. Osiewicz, M., Fache-Rousova, L., & Kulmala, K. M. (2015) Reporting of derivatives transactions in Europe–Exploring the potential of EMIR micro data against the challenges of aggregation across six trade repositories. BIS Report.

40. Perrella, A., & Catz, J. (2020). Integrating microdata for policy needs: the ESCB experience (No. 33). ECB Statistics Paper.

41. RAJU, O. (2021) CREDIT CARD FRAUD DETECTION USING XGBOOST CLASSIFIER.

42. Santos, C. (2013). Bank interest rates on new loans to non-financial corporafions–one first look at a new set of micro data. Economic Bulletin and Financial Stability Report Articles and Banco de Portugal Economic Studies.

43. Sebastian Doerr & Leonardo Gambacorta & José María Serena Garralda, 2021. "Big data and machine learning in central banking," BIS Working Papers 930, Bank for International Settlements.

44. Severino, M. K., & Peng, Y. (2021). Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. Machine Learning with Applications, 100074.

45. Smith, M., Lix, L. M., Azimaee, M., Enns, J. E., Orr, J., Hong, S., & Roos, L. L. (2018). Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy. Journal of the American Medical Informatics Association, 25(3), 224-229.

46. Tam, S. M., & Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. International Statistical Review, 83(3), 436-448.

47. Trisanto, D., Rismawati, N., Mulya, M. F., & Kurniadi, F. I. (2021) Modified Focal Loss in Imbalanced XGBoost for Credit Card Fraud Detection.

48. V. Gudivada, A. Apon, and J. Ding. Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations". In: International Journal on Advances in Software 10.1 (2017), pp. 1 - 20.

49. Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning - a case study of bank loan data. Procedia Computer Science, 174, 141-149.

50. Zambuto, F., Buzzi, M. R., Costanzo, G., Di Lucido, M., La Ganga, B., Maddaloni, P., ... & Svezia, E. (2020). Quality checks on granular banking data: an experimental approach based on machine learning?. Bank of Italy Occasional Paper, (547).

Bank of Russia

MACHINE LEARNING-BASED APPROACHES FOR AUTOMATIC DATA VALIDATION AND OUTLIER CONTROL OF LOAN MICRODATA IN THE BANK OF RUSSIA

Dmitrii Diachkov
STATISTICS DEPARTMENT

Introduction

Goal

Explore the possibility of applying relatively **simple machine learning methods** to improve an **overall quality** of microdata and decisions based on them in the Bank of Russia.

Motivation

Limited human ability (analyst or statistician) to analyze the reliability of a large and complicated datasets that change over time

Field of application

Attributes of loan microdata:

- Borrower’s economic activity;
- Bank’s accounting
- Borrower’s business size;
- Loan type
- Interest rates

Goal achievement criteria

High interpretability of results;

Moderate ease of implementation;

Moderate process control;

High scalability;

High automation capability.

Data sources

- Banks’ reporting form 0409303 "Information on loans granted to legal entities and individual entrepreneurs" with microdata on loans;
- Annual accounting data;
- Statistical Registry on companies provided by State Statistics Service;
- State Registry of SMEs provided by Federal Tax Service.

Stack of technologies – Oracle R Advanced Analytics

Highly reliable organized collection of structured data

RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management

Packages: caret, rpart, neuralnet, nnet, caTools, randomForest, class, cvms

Accumulated experience of working with loan microdata

Quality

- Granular data may be of low quality (Osiewicz et al. 2015) and administrative sources are not exceptions
- Ad-hoc data collection will deliver weak results, which affects the quality, reliability and timeliness of subsequent analysis.(Gross, 2010)
- Matching of micro datasets can be performed to increase data quality (Gadanecz et al. 2016)

High granularity

- The ability to identify objects, operations, contracts, companies etc.
- Presence of outliers that are invisible during aggregation (Santos, 2013)
- Microdata can be joined with other databases data bases allowing an easy identification (Gonzalez et al. 2019)

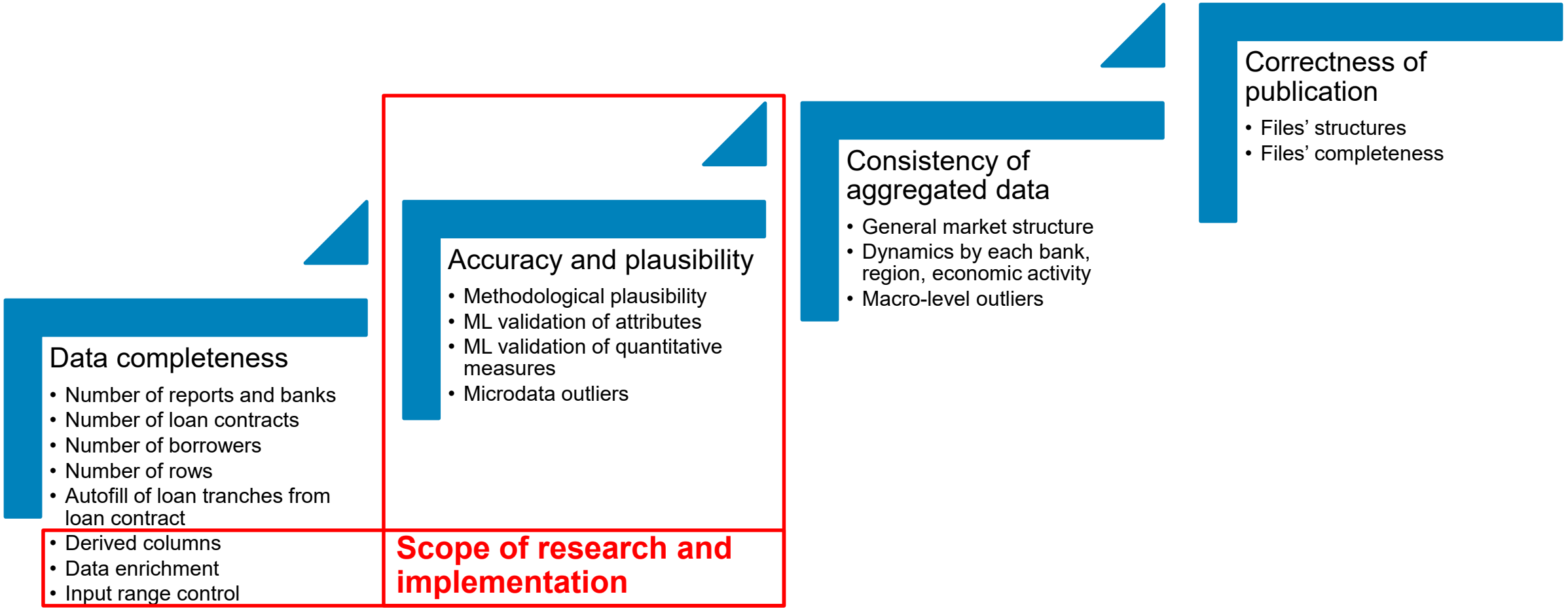
Diversity

- Data values may not conform to allowable (or reasonable) values or ranges
- Efficient use is impossible without full-fledged verification procedures (Crato & Paruolo, 2018)

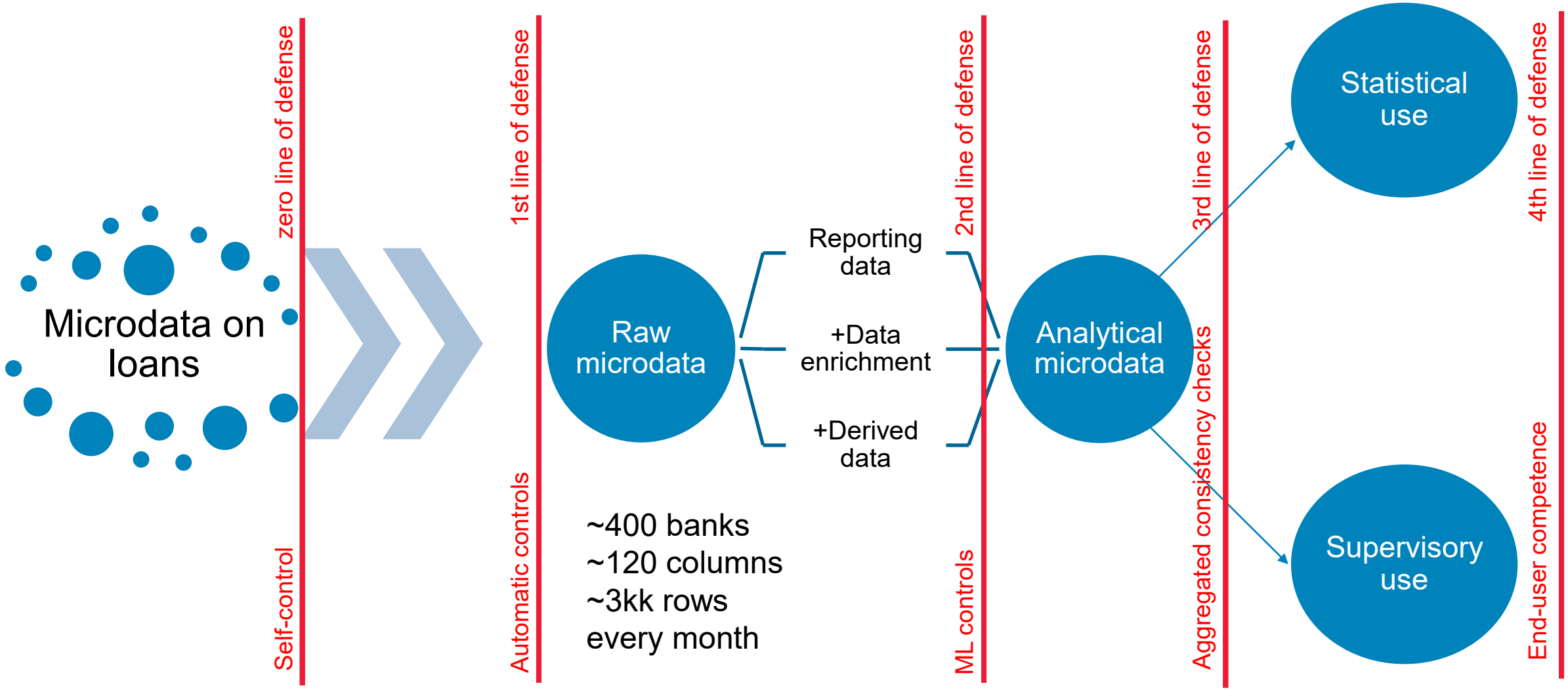
Volumes & Frequency

- The need to efficiently use computing power
- The use of machine learning will reduce the cost of producing statistics and improve its quality (Tam & Clarke, 2015)

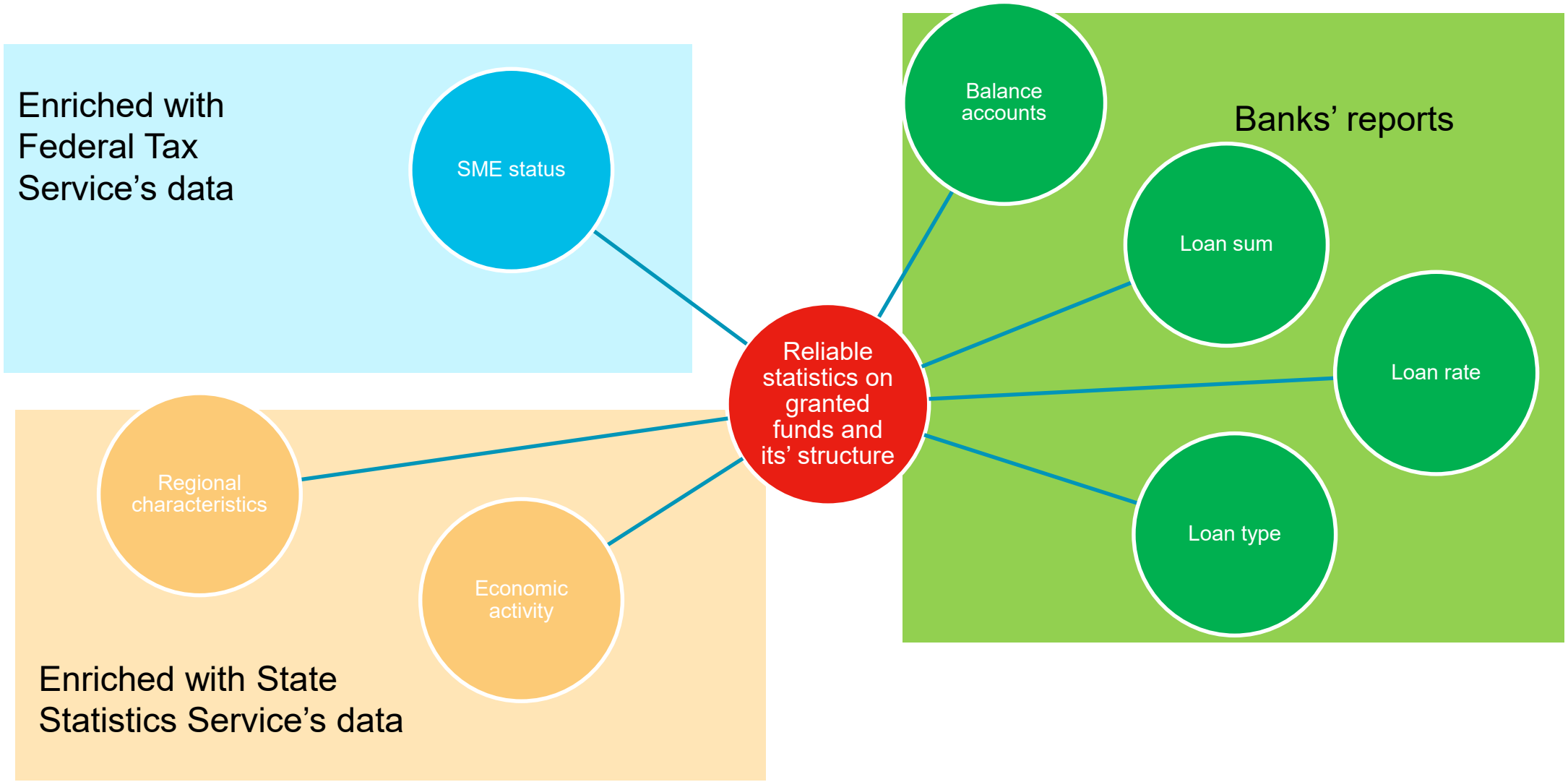
Control procedures for loan microdata in the Bank of Russia



Loan microdata pipeline in the Bank of Russia



Most beneficial ways to apply ML validation (as of today)



Case 1. Validation of balance sheet codes for non-financial companies

Problem

Balance sheet account is a simple and understandable attribute for economists that can be used to filter borrowers by type and industry. Information about the balance sheet account is reported to the Bank of Russia by banks.

Balance sheet code can be determined by bank based on irrelevant data or just by mistake. Another type of mistake is technical errors during report preparation or submission.

Intuition

The balance sheet account should be assigned by bank to it's borrower based on limited list of parameters, such as type of loan, business entity form and economic activity …

…so there must be a way to cross-check balance sheet codes with the data of public registers or other data…

Implementation

A set of decision trees, that establish dependencies between balance sheets codes:

- Loan type
- Economic activity
- Business entity form
- Debt sum

Main balance sheet codes that are used for loan statistics and to be validated:

452 – Loans to non-financial companies

454 – Loans to individual entrepreneurs

451 – Loans to financial companies

Measuring the efficiency of 15 models: balancing accuracy VS speed

TARGET:
BALANCE SHEET CODE == “452”
Train data: 845k loans;
Test data: 211k loans (20%).

The dependent variable is evenly distributed in the training and test samples.

Data type	Yes	No
Train	47,56%	52,44%
Test	47,54%	52,46%

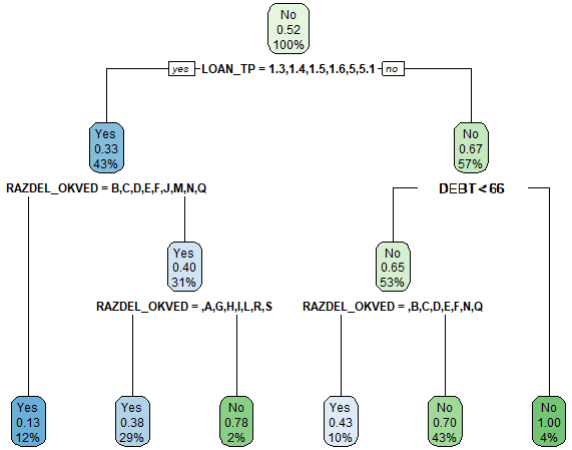
Since the training and test samples are balanced (predictable classes are evenly distributed) -> Accuracy (1) is an excellent metric of model quality.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} (1)$$

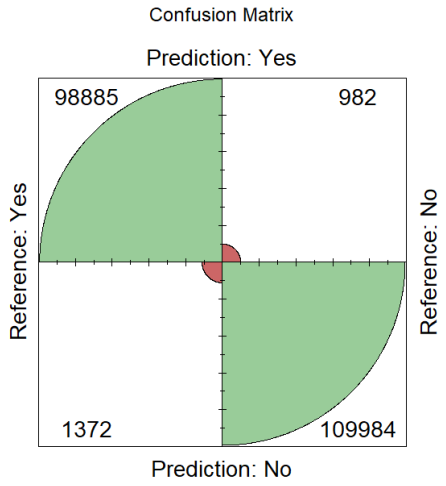
Formula	Mean accuracy, %	Mean time (seconds per 1kk rows)	Relative Efficiency
~ ECON_ACTIVITY + BORROWER_TYPE + LOAN_TYPE	98,904	17,044	5,80
~ ECON_ACTIVITY + LOAN_TYPE + DEBT + BORROWER_TYPE	98,886	18,257	5,42
~ BORROWER_TYPE + LOAN_TYPE	97,352	19,742	4,93
~ BORROWER_TYPE + LOAN_TYPE + DEBT	97,346	20,512	4,75
~ ECON_ACTIVITY + BORROWER_TYPE + DEBT	96,064	22,028	4,36
~ ECON_ACTIVITY + BORROWER_TYPE	95,981	19,546	4,91
~ BORROWER_TYPE	94,182	17,071	5,52
~ BORROWER_TYPE + DEBT	94,155	23,28	4,04
~ ECON_ACTIVITY + LOAN_TYPE	69,813	22,616	3,09
~ ECON_ACTIVITY + LOAN_TYPE + DEBT	69,732	35,284	1,98
~ LOAN_TYPE	67,244	16,122	4,17
~ LOAN_TYPE + DEBT	67,127	30,853	2,18
~ ECON_ACTIVITY	61,317	17,175	3,57
~ ECON_ACTIVITY + DEBT	61,226	32,287	1,90
~ DEBT	54,282	28,043	1,94

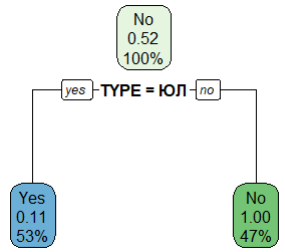
$$Relative\ Efficiency = \frac{E(Accuracy)}{E(Time)} (2),$$

Top-3 models for Case 1: measuring the efficiency

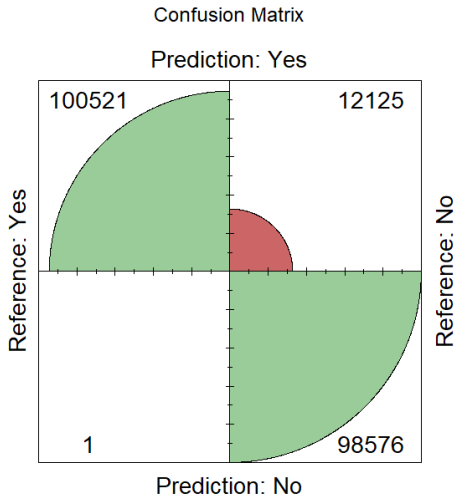


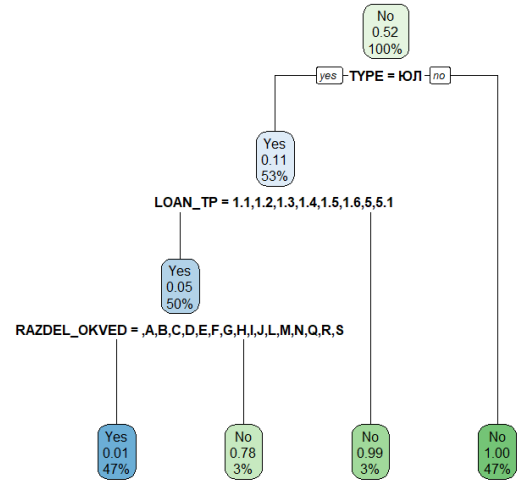
“Accurate”



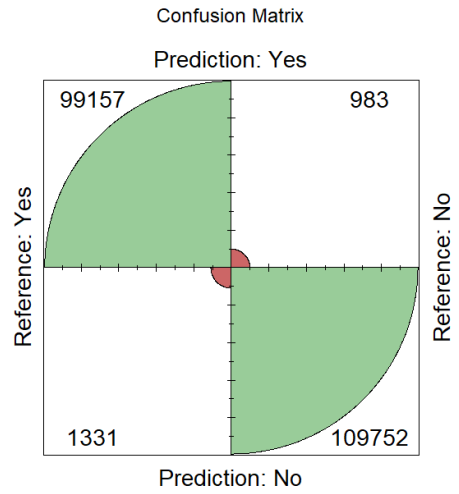


“Fast”





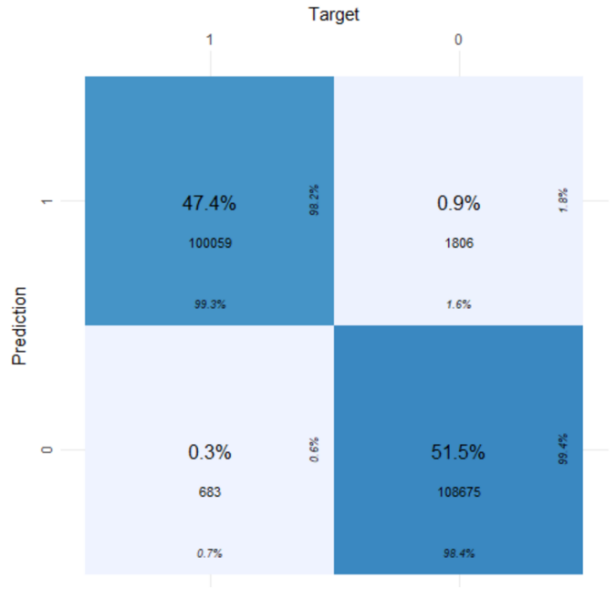
“Balanced”

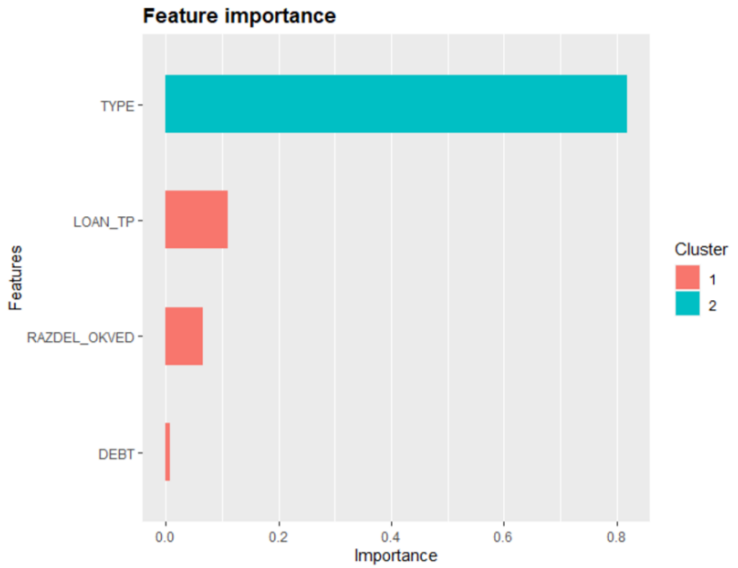


Alternative solutions with XGBoost and logistic regression

We achieved the same results in terms of efficiency but with a longer computational time.

Model type	Mean accuracy	Mean precision	Mean recall	Mean time of fit and predict (per 1kk rows), seconds	Relative efficiency score
Decision tree	0,989	0,987	0,990	17	1,939
XGBoost	0,988	0,993	0,984	45	0,731
Logistic regression	0,989	0,99	0,989	113	0,291





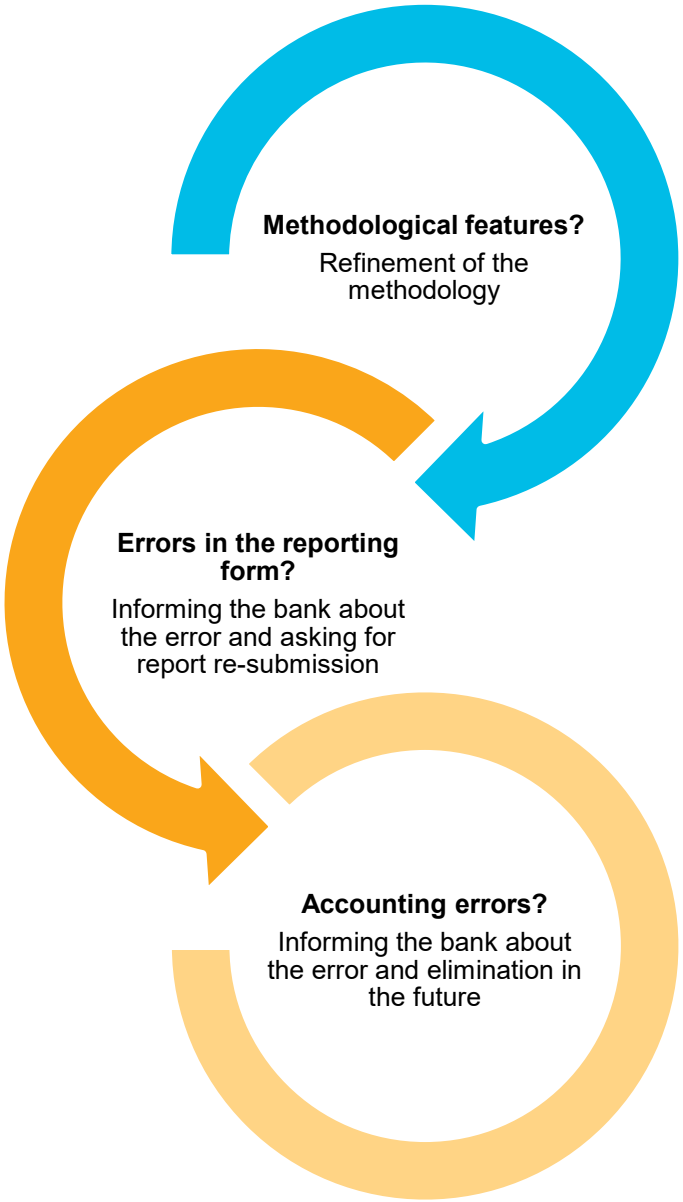
Deviation analysis for Case 1

1

0.66% or 1331 loans were classified by the decision tree as loans of non-financial companies (code 452), but in fact they are reported on other accounts (FP, for example, 453 - loans to non-residents or 451 - loans to financial companies).

2

0.44% or 983 loans were classified by the decision tree as loans that should be recorded as non-452, but in fact they are recorded in 452 accounts (FN, mainly financial companies, which should be reflected with code 451)



Case 2. Balance sheet code 451 with imbalanced distribution

TARGET:
BALANCE SHEET CODE == “451”
Train data: 845k loans;
Test data: 211k loans (20%).

The dependent variable is not evenly distributed in the training and test samples

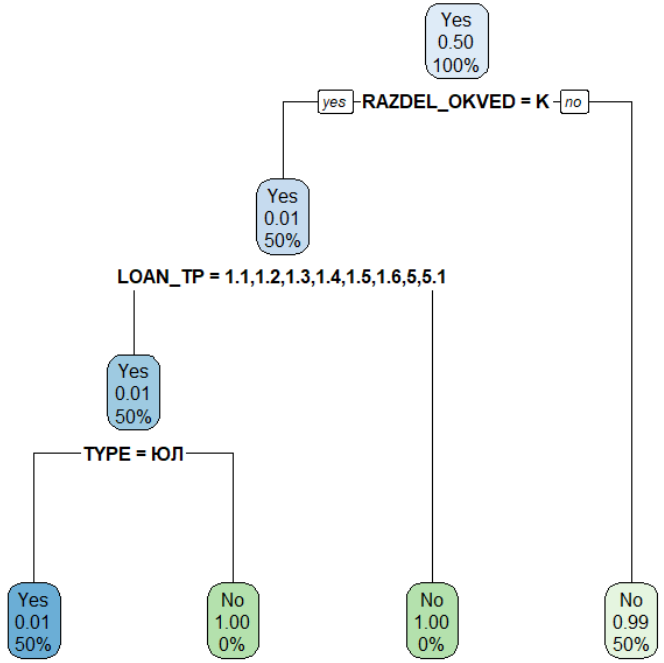
–> imbalanced samples approach

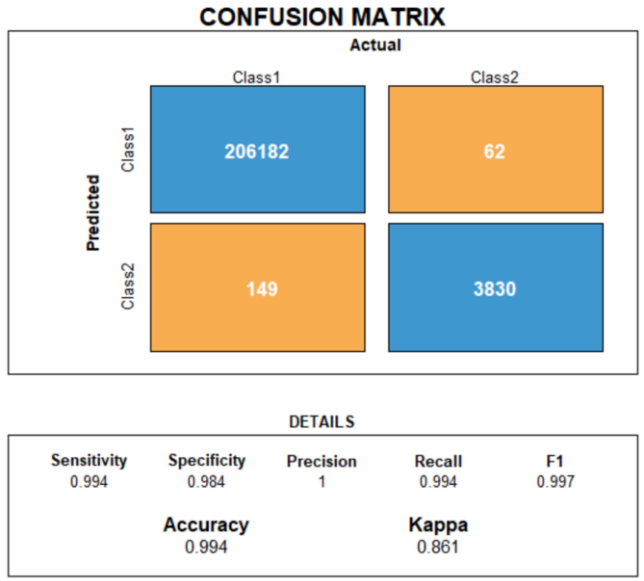
Data type	Yes	No
Train before up-sampling	1,80%	98,20%
Train after up-sampling	50%	50%
Test	1,82%	98,18%

Formula	Mean F1_score, %	Mean time (seconds per 1kk rows)	Relative Efficiency2
~ ECON_ACTIVITY + LOAN_TYPE + DEBT + BORROWER_TYPE	86,465	7,334	11,79
~ ECON_ACTIVITY + BORROWER_TYPE	85,812	6,053	14,18
~ ECON_ACTIVITY + LOAN_TYPE + DEBT	83,952	4,871	17,24
~ ECON_ACTIVITY + LOAN_TYPE	83,718	4,587	18,25
~ ECON_ACTIVITY + BORROWER_TYPE + DEBT	80,405	4,129	19,47
~ ECON_ACTIVITY + BORROWER_TYPE + LOAN_TYPE	80,298	2,916	27,54
~ ECON_ACTIVITY + DEBT	78,346	3,848	20,36
~ ECON_ACTIVITY	77,568	2,731	28,40
~ BORROWER_TYPE + LOAN_TYPE + DEBT	9,313	4,729	1,97
~ BORROWER_TYPE + LOAN_TYPE	9,057	4,157	2,18
~ BORROWER_TYPE + DEBT	6,569	3,753	1,75
~ BORROWER_TYPE	6,565	2,162	3,04
~ LOAN_TYPE + DEBT	4,944	4,052	1,22
~ LOAN_TYPE	4,73	3,662	1,29

$$Relative\ Efficiency2 = \frac{E(F1\_score)}{E(Time)} \quad (3)$$

Adaptation of the approach to the classification of financial companies





$$Precision = \frac{TP}{TP+FP} (4)$$

$$Recall = \frac{TP}{TP+FN} (5)$$

$$F1\_score = \frac{2 * Precision * Recall}{Precision+Recall} (6)$$

The affiliation of the legal entity to the industry K and specific loan types loan both determine the belonging to the code 451.

Despite the high accuracy (99%), the confidence in the model is significantly lower since the training and test samples are unbalanced.

Accuracy in this situation is an incorrect metric.

Precision (4) and Recall (5) metrics do not depend on the ratio of classes and therefore are applicable in conditions of unbalanced samples.

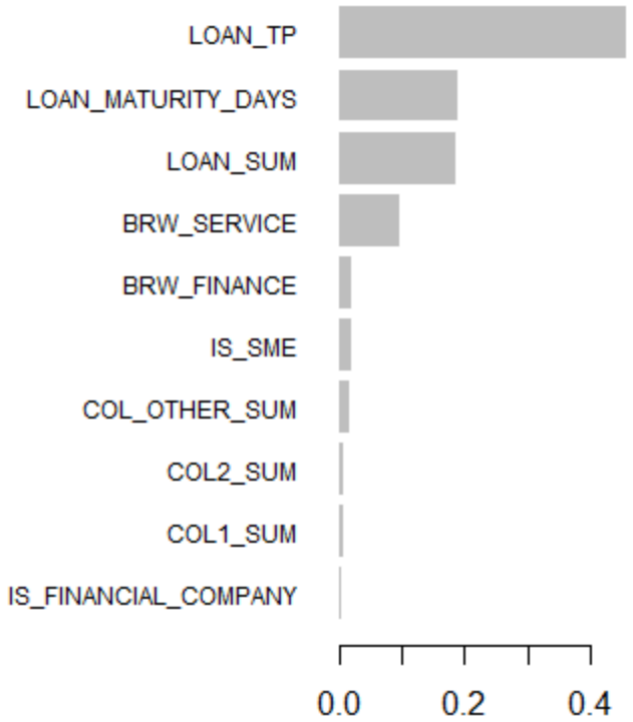
F1_score is a good balance between these two metrics.

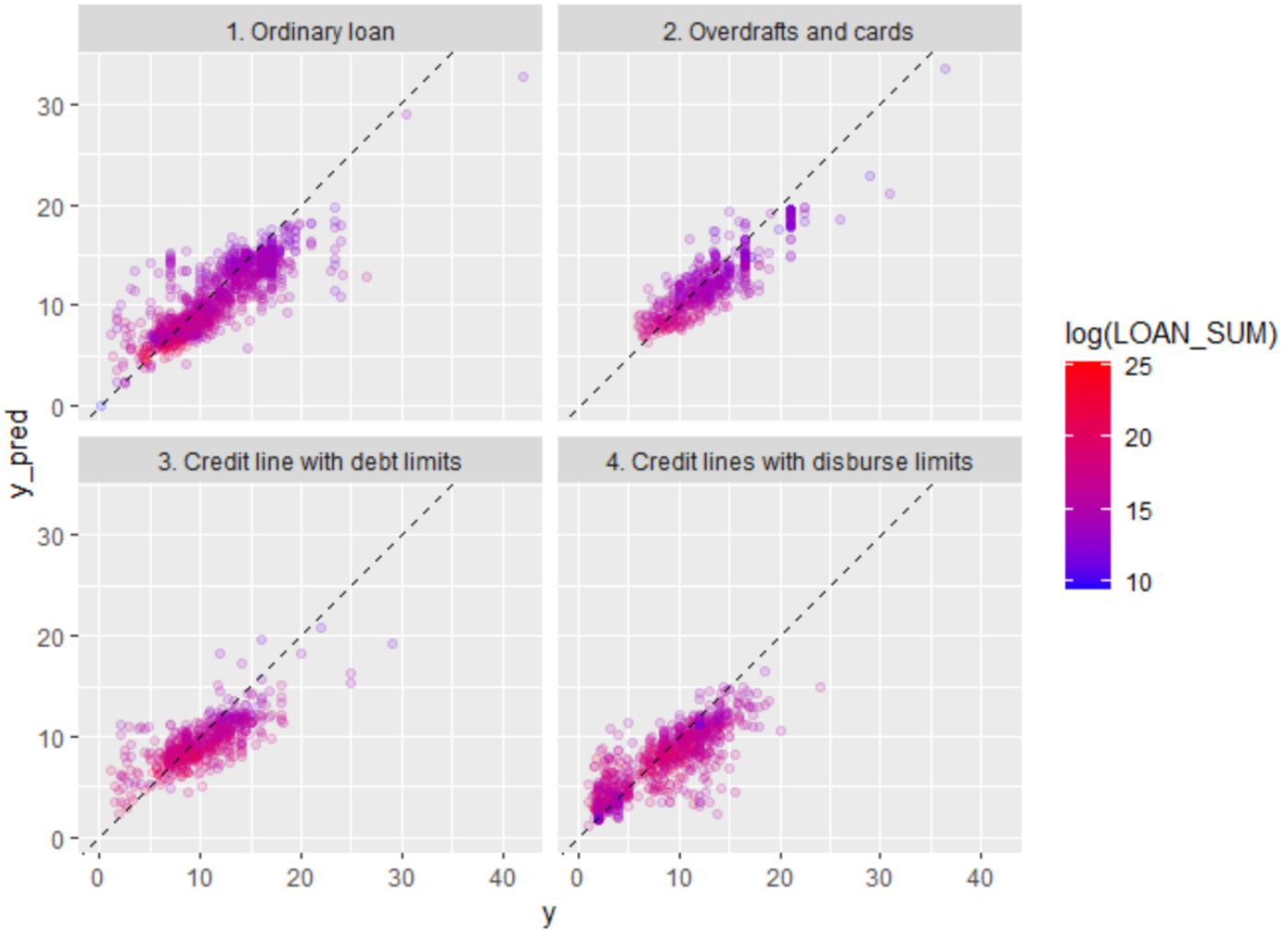
Summary of results for decision tree, XGBoost, and logistic regression approaches to classification task for account code “451”

Train data type	Model type	Accuracy, %	Fit and predict time (seconds per 1kk rows)	F1_score, %	Relative efficiency score 2
Downsampled (30k rows)	Decision tree	98,98%	9,976	99,48%	0,100
	Logistic regression	99,35%	62,34	99,67%	0,016
	XGBoost	99,35%	50,606	99,67%	0,020
Original - imbalanced (845k rows)	Decision tree	99,43%	14,081	99,71%	0,071
	Logistic regression	99,44%	136,62	99,71%	0,007
	XGBoost	99,32%	38,783	99,65%	0,026
Upsampled (1658k rows)	Decision tree	98,99%	15,331	99,48%	0,065
	Logistic regression	99,39%	125,16	99,69%	0,008
	XGBoost	99,35%	70,38	99,67%	0,014

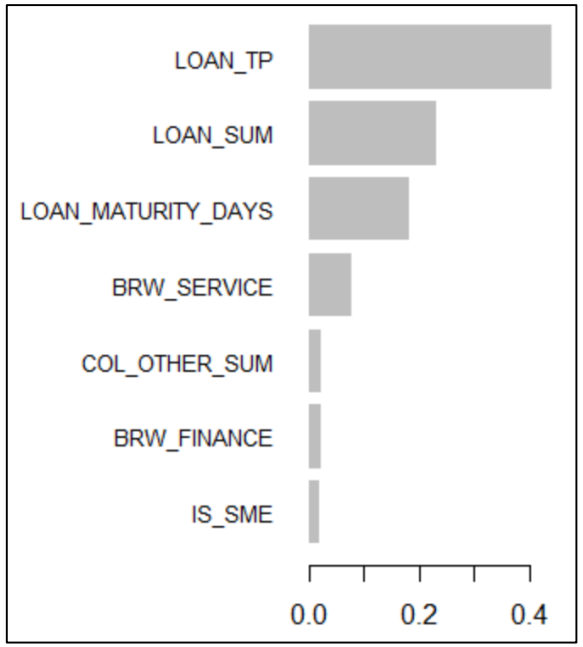
Case 3. Validation of interest rates data with eXtreme gradient boosting

Train data: 421k loans;
Test data: 105k loans (20%).

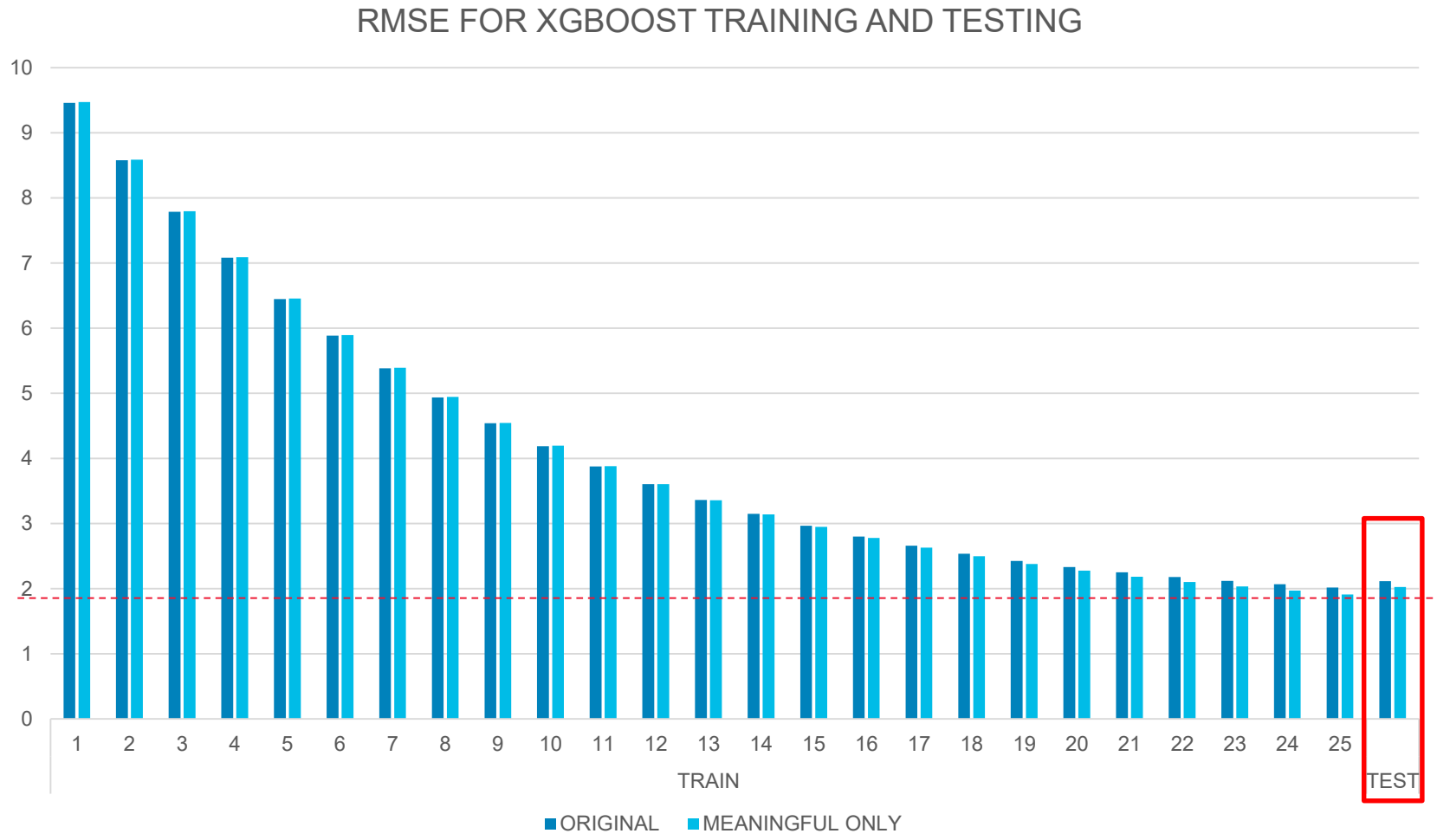




Case 3. Validation of interest rates data with eXtreme gradient boosting

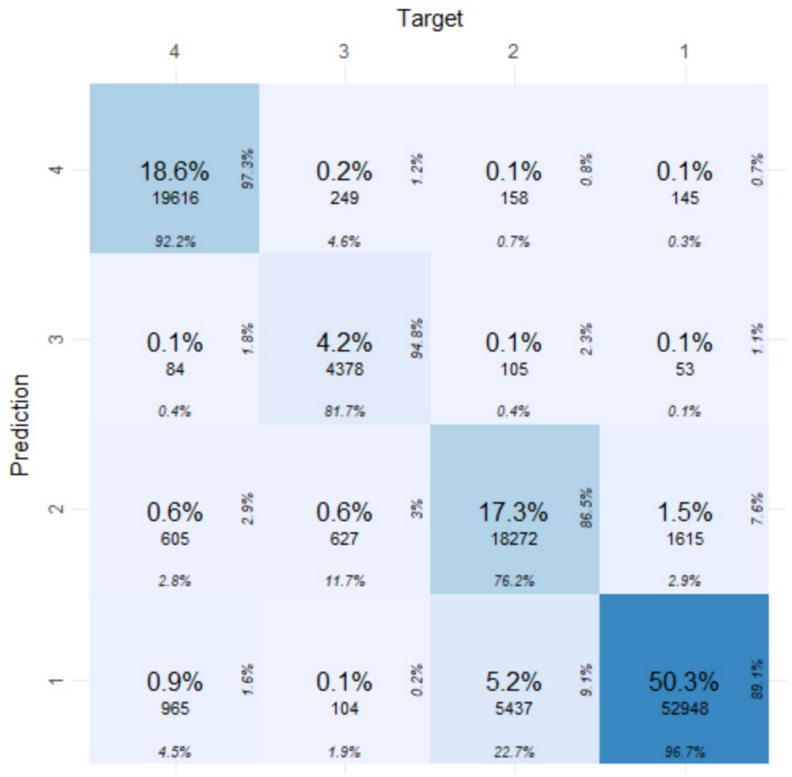


Model	RMSE
XGBoost	2,04
Neural net (caret and nnet)	3,25
Linear Regression	3,85



Case 4. XGBoost multiclassification for SME size validation

Powerful tool with high speed and very high accuracy on millions of rows



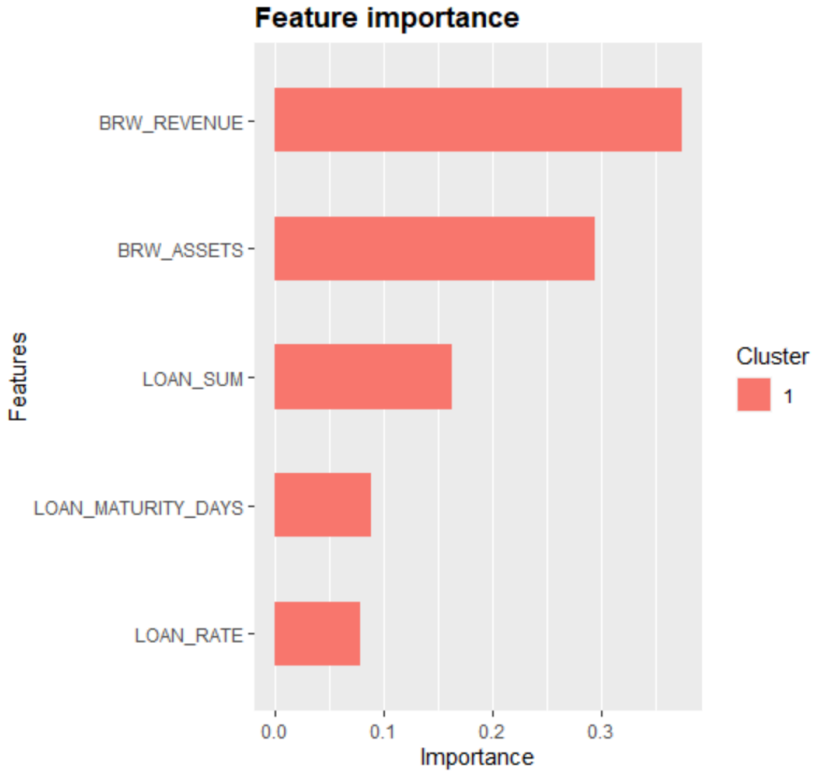
imbalanced

Performance on imbalanced sample is just 2% less accurate, than on upsampled and balanced (90% VS 92%)

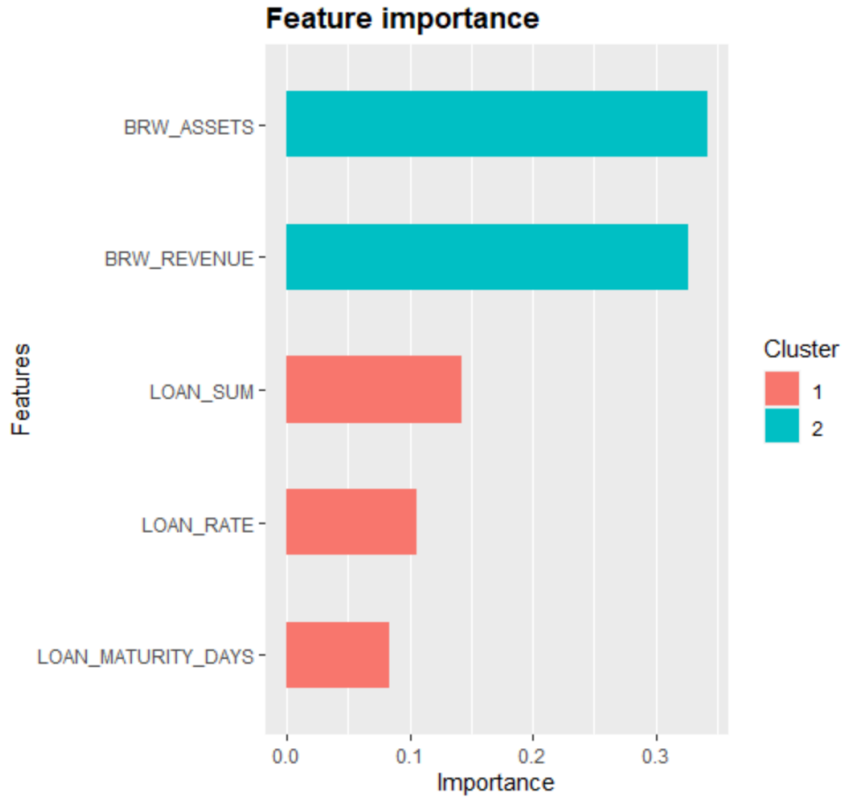
balanced

Case 4. XGBoost multiclassification for SME size validation

But the balanced sample provided more insight on the importance of features, so computational power may be saved



imbalanced



balanced

Case 5. Validation of SME status with neural networks

Problem

Belonging to the SME Registry (provided by Federal Tax Service) defines a borrower's business size. This attribute is extremely important for the analysis of the economic situation. However, really small and insignificant companies may be excluded from the Registry for various reasons. We need to be able to check the validity of any given status.

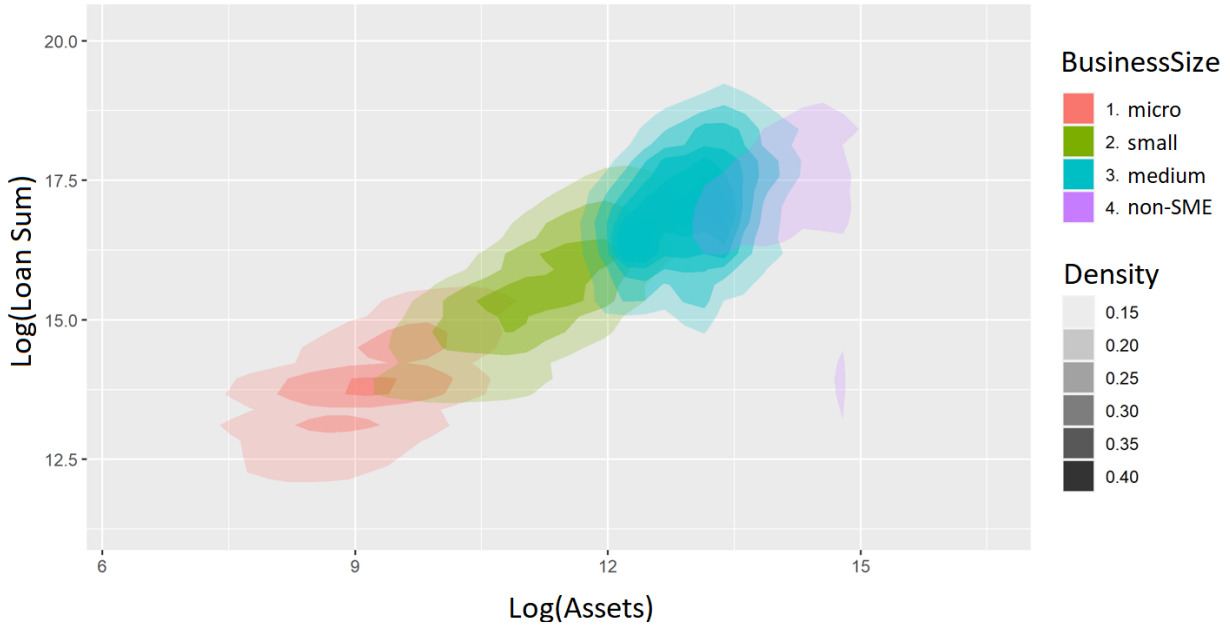
Intuition

Large borrowers have more assets and, accordingly, apply for more significant amounts of loans.

SME borrowers are usually smaller in business size and balance sheet.

Implementation

Build a neural network that can classify companies as SMEs and non-SMEs based on various quantitative indicators characterizing the loan and the borrower.



Finding balance between complexity and accuracy

TARGET:
COMPANY IS SME == “TRUE”

The dependent variable is not evenly distributed in the training and test samples

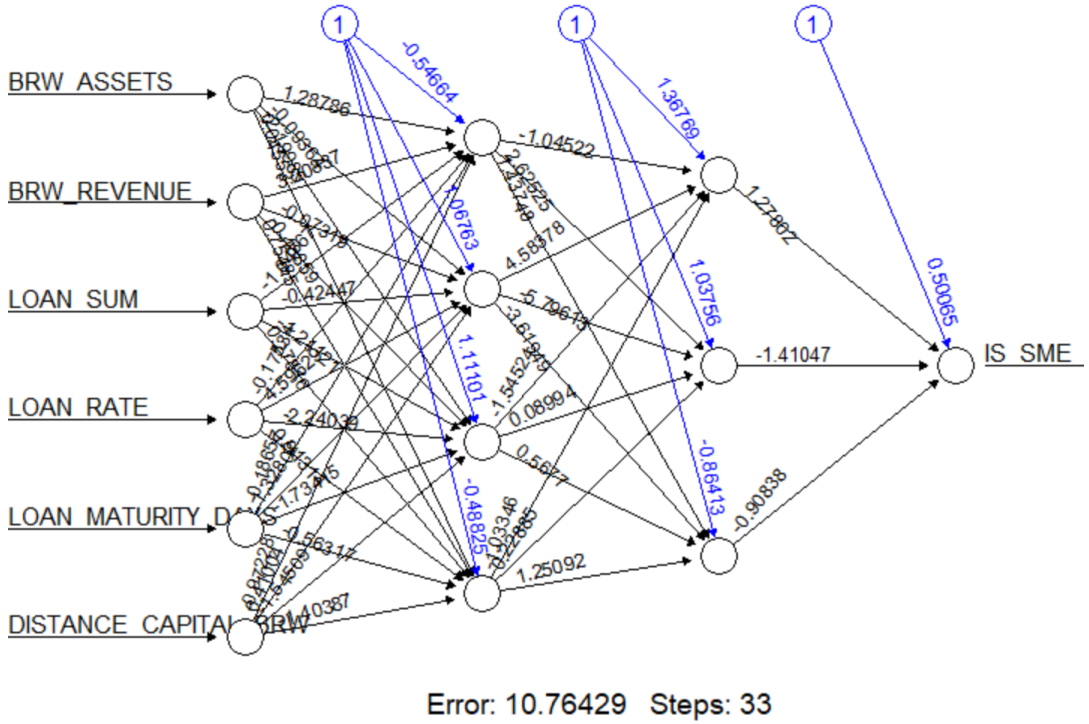
1. Imbalanced samples approach

Data type	Yes	No
Train before up-sampling	75,48%	24,54%
Train after up-sampling	50%	50%
Test	75,09%	24,91%

2. Normalizing the input information for the model

+ scaling or normalization to the range [-1; 1]

+ sigmoid activation function



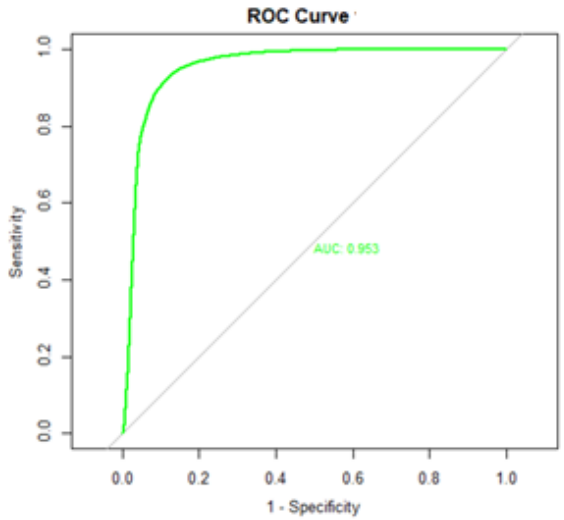
2 hidden layers, 4 and 3 neurons
Accuracy - 90.4%

Training time ~ 619 sec. for 1kk objects

Choice of optimal neural network

Hidden layers	Accuracy	Fitting time	Prediction time	Weighted score
4,3	90,4%	619,78	138,86	10,77211413
2,2	67,2%	102,54	334,39	10,33535241
3,2	73,2%	232,73	299,22	10,07282431
4	79,8%	518,30	245,60	8,336222782
5,3	71,0%	339,18	305,19	7,823119106
4,2	71,2%	264,82	420,88	7,393118623
4,1	80,6%	607,07	346,24	6,814520418
5,3	58,2%	7,68	493,28	6,761507372
3	78,8%	484,72	444,25	6,684266136
4,1	47,4%	5,19	337,11	6,563785262
4,3	55,8%	11,47	539,54	5,650794018
4,2	70,8%	619,94	354,98	5,141586833
2	77,6%	1063,63	156,36	4,935904051
3,2	88,4%	1398,23	281,46	4,652387824
5	69,4%	321,55	760,02	4,453117059
4,3	43,8%	6,85	426,56	4,426365879
4,2	70,6%	796,28	380,96	4,233930891
5,4	76,8%	1431,68	377,27	3,260586787
6,2	77,8%	1619,66	292,87	3,16482278

A simpler neural network sometimes may turn out as more accurate and efficient in terms of consumption of computational resources (when it comes to big data, this issue becomes very significant)



Alternative approach: random forest and logistic regression

Model	Accuracy	F1_score	Computation time (per 1kk rows), seconds
Neural net	90,4%	80,1%	757
Random forest	92,8%	81,4%	318
Log regression	93,3%	83,1%	108

We approached the same task with random forest and logistic regression, and achieved the same accuracy and F1_score results, but faster.

In terms of computational speed, the efficiency of random forest is two times higher, while results are even slightly better.

Traditional classifier logistic regression has beaten neural networks and random forests in terms of result/speed ratio.

Main conclusions

Interpretability, controllability, the possibility of automatic selection of informative features of decision trees, and regressions were the reason for their use as the primary tool for efficient classification of processing large amounts of data, searching for atypical values for subsequent filtering, and identifying erroneous values in categorical variables.

Due to human disabilities, to analyze the reliability of a large and diverse set of data, expanding the field of applied machine learning methods will increase the quality of data and decisions made on their basis.

When solving classification problems, metrics should be monitored carefully and problems solved under business logic.

With unequal classes, metrics should be selected carefully, and up-sampling or down-sampling applied when necessary.

Simpler models often give more balanced and correct results during cross-validation on test data.

Bank of Russia

THANK YOU

Work email: dyachkovdv@cbr.ru
Personal email: d.djachkov@gmail.com