Dimensionality reduction for financial distress detection

AUTHORS:

Dovilė Kuizinienė
dovile.kuizinienė@vdu.lt

Tomas Krilavičius
tomas.krilavicius@vdu.lt

INTRODUCTION

Identification of a business's financial distress is a topic that became relevant together with the birth of money and business. Different methods and tools were used to identify it, and over the last several hundred years different quantitative and statistical methods were getting standard, e. g. for credit scoring. However, with the omnipresence of digital technologies, big data, and Artificial Intelligence, new methods and approaches are being investigated and applied for credit scoring, insolvency, financial distress, and other business indicators analysis and detection.

GOAL

The purpose of this study is to compare different dimensionality reduction techniques for financial distress recognition and important variable extraction.

DATA

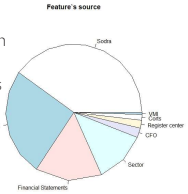
Target: A 'bad' situation of the company, which was detected at least in one Lithuanian government register (indication of True).

Features: 977 different features, including not only financial indicators but also other equally important features like changes in employees, judicial events, managers, etc.

Time: The analysis period is from 2016 to 2022

Sample size: 274105 (bad cases 2.8 %).

Train/Test split: information from last year is included in the testing sample, which makes up 26.3% of the whole sample.



METHODOLOGY

Feature selection techniques:

I group (feature selected before normalization): Correlation, Cohens d, Kruskal;

II group (feature selected after normalization): RF, XGBoost, LASSO;

III group (combination score): overlapping features (the features ranked from all techniques), voted importance (the most ranked features of all techniques).

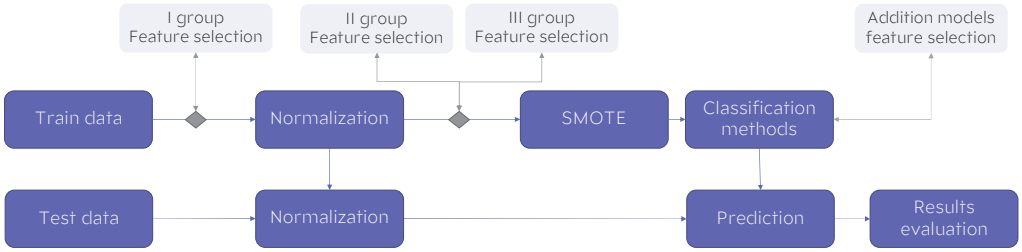
Additional models feature selection: after training classification model the most ranked features were selected, and the model retrained. It was uplaid only on LG, RF and XGBoost models.

Classification models:

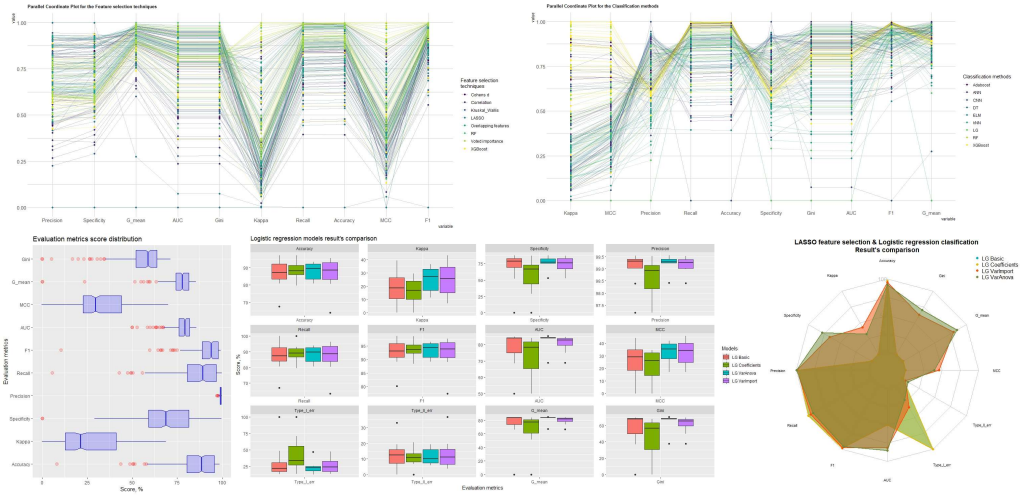
Logistic regression (LG), Decision tree (DT), Random Forest (RF), XGBoost, Adaboost, K nearest neighbors (kNN), Artificial Neural Network (ANN), Extreme Learning Machine (ELM), Convolutional Neural Network (CNN).

Evaluation metrics:

Accuracy, Kappa, Specificity, Precision, Recall, F1-score, AUC, MCC, G-mean, Gini, Error rate, Type I error rate, Type II error rate.



RESULTS



Feature selection technique	Classification model	Accuracy	F1-score	Error rate	Feature selection technique	Classification model	Recall	Type I error rate	Feature selection technique	Classification model	Specificity	Type I error rate	Feature selection technique	Classification model	AUC	Gini
XGBoost	XGBoost Binary	98.56	99.26	1.44	LASSO	Logistic regresion	99.99	0.01	LASSO	CNN 3	99.75	0.25	RF	ANN 1	85.66	71.33
XGBoost	XGBoost Softmax	98.56	99.26	1.44	LASSO	Logistic regresion VarCoef	99.98	0.02	Kruskal Wallis	Decision Tree	93.75	6.25	Voted importance	CNN 5	85.19	70.38
Voted importance	XGBoost Binary	98.55	99.26	1.45	Correlation	CNN 3	99.97	0.03	XGBoost	ELM model 250	93.55	6.45	LASSO	Logistic regresion VarAnova	85.13	70.27
Voted importance	XGBoost Softmax	98.5	99.23	1.5	XGBoost	XGBoost Softmax	99.74	0.26	Cohens d	CNN 3	92.66	7.34	Kruskal Wallis	CNN 2	84.9	69.79
XGBoost	Random Forest	98.46	99.21	1.54	XGBoost	XGBoost Binary	99.73	0.27	Overlapping features	ANN 3	92.42	7.58	XGBoost	Logistic regresion	84.74	69.47
RF	Random Forest	98.42	99.19	1.58	Voted importance	XGBoost Binary	99.72	0.28	LASSO	ANN 3	92.22	7.78	RF	Logistic regresion VarImp	84.67	69.35
Voted importance	Random Forest	98.32	99.14	1.68	Voted importance	XGBoost Softmax	99.66	0.34	Kruskal Wallis	ANN 3	92.22	7.78	Voted importance	Logistic regresion	84.66	69.32
XGBoost	XGBoost less depth Binary	98.29	99.12	1.71	XGBoost	Random Forest	99.61	0.39	LASSO	ANN 1	92.17	7.83	RF	ANN 3	84.6	69.19
XGBoost	XGBoost less features Softmax	98.24	99.1	1.76	RF	Random Forest	99.57	0.43	Overlapping features	ELM model 150	91.97	8.03	Overlapping features	CNN 3	84.57	69.15
Voted importance	XGBoost less features Binary	98.19	99.07	1.81	XGBoost	XGBoost less depth (Binary)	99.48	0.52	Cohens d	CNN 2	91.63	8.37	XGBoost	Logistic regresion VarAnova	84.38	68.76

CONCLUSIONS

- XGBoost performed similar results irrespective of feature selection technique selection.
- LG models results improved using aditional dimensionality reduction technique after models trainig (VarAnova or VarImport).
- 204 experiments were conducted, based on which different combinations of the best methods were determined in order to achieve the highest accuracy (XGBoost-Xgboost), recall (LASSO-LG), sspecificity (LASSO-CNN3), AUC (RF-ANN1).

FUTURE PLANS

- Using feature extraction methods: PCA, LDA, t-SNE.
- Using unsupervesed machine learning methods: Isolation forest, K-means, One-Class SVM (OCSVM).
- Adding additional information about bad CFO history cases.
- Focusing on other target group determination including more insolvency class indications.

CARD
CENTRE FOR APPLIED RESEARCH AND DEVELOPMENT