Croatian Operational Research Review
CRORR **16:2**(2025), 165–176

Going concern prediction – A horse race between traditional and regularization machine learning models

Tina Vuko[1,*], **Slavko Šodan**[2] **and Ivana Perica**[2]

[1] *Faculty of Economics, Business and Tourism, University of Split, Cvite Fiskovića 5, 21000 Split, Croatia*
E-mail: ⟨tina.vuko@efst.hr⟩

[2] *Faculty of Economics, Business and Tourism, University of Split, Cvite Fiskovića 5, 21000 Split, Croatia*
E-mail: ⟨{ssodan, iperica}@efst.hr⟩

Abstract. Regularization machine learning (ML) methods have been increasingly applied in accounting research, offering new possibilities in predictive modeling. Their forte lies in the effective regularization methods for resolving the biggest concern of generalization, which is the risk of overfitting the training data. While these sophisticated methods are known to outperform traditional regression approaches in large and balanced datasets, this may not be the case when facing imbalanced and small datasets. Moreover, model validation is also challenging in such settings because traditional performance measures, such as prediction accuracy, may be misleading. We address this problem by comparing two traditional and five regularization-based methods in predicting going concern uncertainty (GCU) on the sample of listed companies in Croatia. We take caution when evaluating the models due to class-imbalanced problems and include different classification performance measures, as well as calibration of the models to account for their uncertainty. As expected, no model performs best across all evaluation criteria, but regularization methods are better calibrated. Given our results, we suggest that model selection should consider the results of the model calibration, a combination of different performance metrics, and the economic impact of the statistical performance of the model, if feasible.

Keywords: elastic net, lasso, ridge regression, class imbalance, prediction

Received: October 31, 2024; accepted: December 17, 2024; available online: April 2, 2025

DOI: 10.17535/crorr.2025.0014

Original scientific paper.

1. Introduction

Going concern (GC) is the key assumption in financial reporting, which assumes that management has no intention nor will be forced to terminate operations and liquidate a company's assets at least within the next 12 months. However, when there is substantial uncertainty about a company's ability to continue as a GC, an auditor should emphasize this in the audit report. Therefore, GC emphasis in the auditor report is the result of an auditor's prediction of a GCU for a given company, taking into account all available information for the near future. The auditor's GC emphasis should be understood as an early warning sign of the potential inability of the company to continue as a GC and not as an absolute and ongoing assurance of

*Corresponding author.

the company's ability to withstand future shocks. Since GCU is a complex and highly judgmental decision with widespread consequences [8], it is interesting from the aspect of predictive modeling.

One of the major problems when predicting GCUs (and many other interesting accounting events, such as fraud, accounting restatement, or bankruptcy) is that they occur infrequently. Consequently, researchers face rather imbalanced datasets and relatively small sample sizes compared to the number of potential predictors. Building a predictive model in such circumstances can be a significant challenge both from the aspect of selecting important variables and from the aspect of model evaluation and selection (i.e. choosing the final model among a set of models). Over the years, logistic regression (LR) has become the most widely used method to predict discrete outcomes in accounting research [20]. Likewise, LR is the most frequently used statistical method to predict auditors' GCU decisions [5], [21], [8]. While LR has many attractive properties, a potential limitation from the aspect of predictive modeling is related to variable selection. Namely, traditional variable selection methods like stepwise selection may lead to overfitting. Also, due to pre-testing bias, they may overlook some important but insignificant predictors, leading to the loss of predictive information [16]. This limitation is strongly emphasized when dealing with small and imbalanced datasets and when it is not apparent whether a variable should be included in the model or not.

Recent advancements in statistical and ML algorithms offer new opportunities for variable or feature selection and prediction problems. Shrinkage methods such as *lasso* (the least absolute shrinkage and selection operator) are known to address some of these challenges through regularization techniques. The attractiveness of shrinkage methods in predictive learning models is majorly based on the bias-variance tradeoff, following the general idea that prediction accuracy can be improved by shrinking the values of the regression coefficients or setting some coefficients to zero [16], [17]. However, the shrinkage does not necessarily lead to better predictive performance, particularly with small sample sizes. Moreover, when the sample size is small, it may be even advisable not to build a prediction model [28]. Given these concerns, we question if sophisticated statistical and ML algorithms that have been increasingly applied in accounting research [6], [14], [20] can outperform traditional statistical methods in predictive modeling of rare classes.

To answer this question, we illustratively use auditors' GCU predictions from a sample of companies listed in Croatia over ten years. Our dataset can be described as small and moderately imbalanced. In order to build competing models, we define 25 plausible covariates. We focus on predictors that can be easily calculated from publicly available financial statements and are commonly used to evaluate a GCU (see for example Cyhe Koh et al., 2004; Martens et al., 2008; Yeh et al., 2014; Goo et al., 2016 studies [9], [21], [31], [15]). In addition to financial indicators, we use three audit-related indicators: type of auditor (Big Four auditors or others), type of audit opinion (unmodified audit opinion or modified audit opinion), and change of auditor (a company has changed auditor or not). Since they do not apply any regularization, we use conventional LR with stepwise variable selection (*LRsw*) and a full set of predictors (*LRfull*) as the reference methods. Finally, we compare the predictive performance of standard LR models with five regularization-based models (*lasso, adaptive lasso, plugin lasso, ridge, and elastic net*) to assess the predictive performance of the models. Given that our sample size is relatively small and that some covariates are highly correlated, there is a substantial chance of overfitting.

An additional problem arises when evaluating models with unequally distributed dichotomous outcomes. Commonly used model evaluation metrics such as prediction accuracy, an overall error rate, or a proportion of correctly (miss)classified samples might be misleading as they favor classifiers that accurately predict the majority class [7]. On the other hand, class balancing by applying different sampling techniques to overcome imbalanced data structure is problematic because it makes the number of minority and majority classes predetermined and

not random [7]. To overcome this problem, we use a wide range of classification metrics and calibration belts to evaluate the competing models' predictive power comprehensively.

In a way, our research extends Bertomeu's (2020) and Krupa and Minutti-Mezza's (2022) studies ([6], [20]), which provide guidance on conducting accounting-related research involving models that predict discrete outcomes using traditional and ML algorithms. However, we focus on methodological challenges related to prediction modeling and valuation when facing imbalanced and small datasets. Our results indicate that while nowadays it is fashionable to construct different models using sophisticated statistical and ML methods to predict important accounting outcomes, more effort should be put into discussing how and why such models work or what underlying issues they address [29]. Different research design choices (variables selection, splitting methods, sampling strategies, model evaluation metrics, etc.) may obscure the comparison between the alternative prediction models. For example, when evaluating GCU predictive models, researchers usually report only a few classification measures, mainly prediction accuracy, ROC area, Type I and Type II errors, precision and recall rates [9], [21], [31], [15], [18], [23], [24]. Krupa and Minutti-Mezza [20] report similar findings for other mainstream accounting research on predictive modeling. In addition, studies that predict dichotomous accounting events rarely (if any) assess the calibration of the outcome models. We find this important as our results indicate that when predicting infrequent outcomes, no model performs best across all evaluation criteria, or moreover, several models perform about equally well using different algorithms and selecting different variables. Consequently, we may easily select the poorly calibrated model even when considering class-imbalance-adjusted performance measures.

The paper proceeds as follows. Section 2 compares traditional classification methods vs. regularization-based ML methods. Section 3 describes research procedures and the sample characteristics, while section 4 presents research findings. Section 5 concludes with a discussion of the results, limitations of our approach, and potential implications for future research.

2. Traditional vs. regularization methods

2.1. Variable selection

Over the years, LR has become a popular prediction model in accounting because it produces a linear combination of the variables with weights and confidence intervals for the weights that clearly show how the predictors affect the outcome variable. However, when the sample size is small or imbalanced, traditional classification methods could produce overfitted and unstable models, making model selection problematic. Namely, a stepwise regression approach to variable selection suffers from overfitting and pretesting biases [1], [16].

Therefore, regularization has become the cornerstone of modern statistics. The major benefit of regularization or shrinkage methods is that they can accommodate large predictor models (usually under the assumption of sparsity) and rely on tuning parameters. Regularization methods solve an optimization problem using two terms: fit measure (e.g. deviance ratio as a value of loss function) and penalty term (regularization parameter), also called tuning parameter. The term is included to penalize the complexity of the model. The lasso algorithm penalizes the absolute size of coefficients (L1 or lasso regularization), the ridge penalizes the sum of squared coefficients (L1 or ridge regularization), while the elastic net applies a mix of lasso and ridge-type penalties [1]. Therefore, contrary to the lasso which yields sparse solutions by setting some coefficient estimates precisely equal to zero, the ridge method preserves all predictors in the model [2]. Lasso commonly selects penalty term through cross-validation (CV), the adaptive and plugin methods. Adaptive lasso uses adaptive penalty weights for the lasso penalty term (L1) to achieve the oracle properties of the estimators [32]. Plugin lasso iteratively calculates the smallest penalty term that is large enough to dominate the estimation error in the coefficients. While plugin includes covariates in the model that best approximates

the data, it risks missing some covariates with large coefficients and finding only some with small coefficients [3], [4]. Since the estimates' magnitudes are somewhat restricted, the regularized regression estimator will tend to have a smaller variance than traditional regression methods. Consequently, they may deliver an overall more accurate prediction, particularly in the presence of small samples [27]. Regularization methods also include nonlinear terms and interactions among the predictors without explicitly specifying them [20]. Furthermore, due to their low variance, they are immune to multicollinearity and robust to high-dimensional settings (i.e. the course of dimensionality).

On the other hand, regularization methods trade-off between estimator bias and variance so that they may underestimate, to some extent, the population parameters and fail to capture essential regularities in the training data. Another problem could arise if regularization is applied to domains in which sparsity is not a plausible assumption, and this is particularly questionable when the sample size is small relative to the number of the model's parameters [13]. While ridge regression shrinks the coefficients of correlated predictors towards each other, lasso tends to select one of them somewhat arbitrarily, so one may erroneously conclude that the selected covariate is essential, even though other more important covariate may be left out.

The recent success of some overparametrized models (e.g. neural networks) also indicated that in some cases, very complex models predict better [11] [13] [30]. However, such complex models are hard to interpret and tend to be unstable, so simple, easily interpretable models like LR might often perform just as efficiently as complex ML models [13]. Most importantly, when using regularization methods to develop prediction models on small samples or when having a larger number of potential predictors relative to the number of minority classes, researchers advise more caution and call for more research investigating the impact of specific combinations of shrinkage and tuning methods [26], [28].

2.2. Evaluation of the prediction models' performance

Successful ML training approaches often rely on sufficiently large and balanced data. However, many important real-world events generate only imbalanced, small data sets and pose problems for both traditional and sophisticated statistical or ML algorithms. Commonly used model evaluation criteria such as accuracy, ROC, or ROC AUC metrics can be misleading in this case since statistical and ML models are generally designed around the assumption of balanced class distribution. The high accuracy of such models might reveal more about the underlying distribution of classes than about actual model performance. In a situation like GCU prediction, where false negatives (FN) incur greater cost than false positives (FP), imbalance may lead to adverse consequences. Nevertheless, adding a specific cost-sensitive model to the training data may induce bias in the model if the true error cost of the minority class differs from that of the training data [29]. Also, frequently used sampling techniques (matched sample designs, oversampling the minority class or undersampling the majority class) on training samples are not a preferable solution to the class imbalance problem as they alter the relationship between majority and minority classes, which may affect the incidence [7], [21]. Therefore, the critical issue when evaluating alternative prediction models is the choice of an appropriate performance measure. However, finding the most appropriate and meaningful evaluation metrics for imbalanced data is not achievable without having accurate cost information which could utilize cost-sensitive learning to produce an accurate classifier [29].

To overcome this problem, we use different discrimination and classification metrics to select the model that yields good results over a wide variety of assumptions. In addition to the ROC curve that is commonly used to evaluate the performance of prediction models, we use the precision-recall curve that is better suited for imbalanced datasets [10], as well as different metrics derived from the confusion matrix. Using the confusion matrix, we calculate typical performance measures (accuracy, recall or sensitivity, specificity, precision or positive predicted

values and negative predicted value) and several imbalance-adjusted measures (Balanced accuracy, F-measure, Geometric mean and Matthew's Correlation Coefficient). However, even when the model has good discrimination, the estimated risks can be unreliable if the model is poorly calibrated [28]. Therefore, we use calibration belts to examine the relationship between estimated probabilities and observed outcome rates.

3. Procedure and sample description

To address our research question, we use hand-collected data from financial statements and related audit reports of Croatian listed companies, excluding companies in the financial sector. The sample period covers ten years, from 2009 to 2018. After controlling for the missing data, the initial sample of 929 company-year observations is reduced to 891 company-year observations. For this analysis, we use all years affected by the GCU as our positive classes, so the proportion of positive classes (GCUs) is 18.52%. We define 22 financial and three audit-related variables as possible and commonly used GCU predictors (see Appendix A for variable description and descriptive statistics, available at: Appendix.pdf). We use Stata 18 software for our analyses.

Since our main objective is to compare the performance of traditional LR and regularization prediction models when classes are not equally represented in the dataset, we do not use a pre-specified model nor focus on any specific variable as a potential determinant of GCU by trying to answer whether it reflects an independent mechanism of the outcome. Rather, we focus on assessing the performance of example models when variable selection is used to specify a model using a variety of classification measures and calibrations. Namely, we apply different types of regularization (*lasso*, *adaptive lasso*, *plugin lasso*, *elastic net*, and *ridge regression*) to develop prediction models. We use penalized coefficients estimated on the training dataset to make predictions for regularization methods and unpenalized coefficients estimated on the training dataset to make predictions for traditional LR models. Traditional LR models are developed using stepwise selection (*LRsw*) and a full set of predictors (*LRfull*) to understand better the risk of overfitting on model performance. Because there is a certain trade-off between using fewer data to train (i.e. parameters will have greater variance) and using fewer data to test models (i.e. performance statistics will have greater variance), researchers can use different sample split strategies (for example see Kalinic Milicevic and Marasovic [19]). Therefore, we first assign 2/3 of the sample to the training and 1/3 to the test sample (67:33% sample split). Alternatively, we use an 80:20% sample split as a robustness check. We also set the random-number seed option so that we can reproduce our results.

In the case of ML algorithms, the training sample also uses sample split referred to as the "k-fold cross-validation" where the dataset is divided into k equally sized subsets. One subset is chosen to be the validation set and the remaining *k-1* folds are used to train the model. The validation set is then evaluated with a performance metric such as the deviance ratio. This method is repeated k times so that each subset acts as a validation set exactly once. The performance metric is then averaged across all k iterations to give an estimated performance for the model [12]. Table 1 presents the list of predictors selected by shrinkage methods.

As expected, the results from Table 1 show that the plugin produces the most parsimonious model (only four predictors). The most important variables selected by the *plugin* are profit/loss indicator (*x17*), working capital ratio (*x04*), return on equity (*x15*), and leverage ratio (*x05*). Lasso selected ten predictors; *adaptive* lasso selected nine, and *elastic net* selected 13 predictors, which is the same number of predictors selected by the stepwise method (*LRsw*) (untabulated).

The results of an alternative sample split can be found in Appendix B (available at: Appendix.pdf).

Variable	lasso	adaptive	plugin	elastic net	ridge
cons	x	x	x	x	x
x17	x	x	x	x	x
x04	x	x	x	x	x
x24	x	x		x	x
x15	x	x	x	x	x
x12	x	x		x	x
x05	x	x	x	x	x
x21	x	x		x	x
x18	x	x		x	x
x16	x	x		x	x
x10	x			x	x
x14				x	x
x13				x	x
x19				x	x
x20					x
x07					x
x23					x
x02					x
x11					x
x01					x
x25					x
x03					x
x09					x
x06					x
x22					x
x08					x

Table 1: Standardized coefficients sorted

4. Results

We evaluate the performance of the two traditional LR and five regularization-based GCU prediction models on the hold-out sample using several metrics. We start our analysis by comparing deviance (D) and deviance ratio (D^2) goodness-of-fit statistics, principally used for generalized linear models (GLM). Smaller D and larger *D2* indicate a better model (Table 2).

When looking at the test sample results, D^2 shows that ridge regression performs as the best model and as the most robust model relative to the alternative sample split strategy. We can notice that a simpler model with stepwise selection (*LRsw*) performs better on the test sample data than the model without variable selection (*LRfull*).We can also notice that traditional LR models generally perform better than regularization models on the training sample. In contrast, regularization models perform as effectively or slightly better than traditional LR models on the test samples. This is expected since traditional estimation methods are based on optimizing the estimated model's in-sample fit without any regularization to optimize the out-of-sample fit. However, plugin lasso performs as the worst model. As already explained, the plugin is a rigorous lasso that tends to favor very parsimonious models by selecting the covariates that best approximate the data but also runs the risk of missing some important covariates.

Given that in the test sample deviance ratio is the same or similar for alternative models (e.g. *LRsw* and *adaptive* lasso or *lasso* and *elastic net*), we can conclude that the importance of selected predictors can vary between the models having about the same D^2.
While we can opt for a simpler prediction model in this case, the question remains whether

Model	Sample	Deviance	Deviance ratio
LRSW	1	0.639	0.340
LRFULL	1	0.614	0.365
lasso	1	0.693	0.284
adaptive	1	0.656	0.322
plugin	1	0.826	0.147
elastic net	1	0.696	0.281
ridge	1	0.715	0.261
LRSW	2	0.712	0.241
LRFULL	2	0.747	0.204
lasso	2	0.707	0.247
adaptive	2	0.712	0.241
plugin	2	0.795	0.153
elastic net	2	0.706	0.247
ridge	2	0.702	**0.251**

Table 2: Goodness of fit (Deviance ratio)

Notes: *Sample 1 is a training sample n=594 and Sample 2 is a test sample n=297. The highest D^2 values in the test sample are in bold.*

those models are equally good in prediction accuracy and stability.

In order to evaluate the classification performance of alternative models, we use results obtained from a confusion matrix. Confusion matrix commonly uses a 50% decision threshold so that all values equal or greater than the threshold are assigned to one class and all other values to another. Using the information provided by the confusion matrix, we calculate some commonly used and several imbalance-adjusted classification metrics (Appendix C provides a detailed description of each measure used, available at: Appendix.pdf).

Evaluation metrics	LRSW	LRFULL	Lasso	Adaptive	Plugin	Elastic Net	Ridge
Accuracy (%)	**87.2**	86.5	86.2	**87.2**	82.5	86.2	86.2
Recall (%)	43.4	**47.2**	30.2	45.3	3.8	30.2	28.3
Specificity (%)	96.7	95.1	98.4	96.3	**99.6**	98.4	98.8
Precision (%)	74.2	67.6	80.0	72.7	66.7	80.0	**83.3**
NPV (%)	88.7	**89.2**	86.6	89.0	82.7	86.6	86.4
F-measure	0.55	**0.56**	0.44	**0.56**	0.07	0.44	0.42
F-adjusted	**0.65**	0.62	0.60	**0.65**	0.15	0.60	0.60
BA	0.70	**0.71**	0.64	**0.71**	0.52	0.64	0.64
MCC	0.50	0.49	0.44	**0.51**	0.13	0.44	0.43
GM	0.65	**0.67**	0.54	**0.66**	0.19	0.54	0.53

Table 3: Confusion matrix (test sample results)

Notes: *The highest values of each evaluation metric are in bold.*

Table 3 shows that all models achieve overall accuracy above 82.5 % (the bottom value is for the *plugin*) but not higher than 87.2% (*LRsw* and *adaptive*). Although no model outperforms other models across all metrics when looking at the overall number of the highest values of specified evaluation metrics, *LRfull*, and *adaptive* lasso seem to perform as the best models. All classifiers perform quite similarly concerning accuracy, specificity, and negative predicted value (NPV).

The difference between classifiers is more pronounced when observing precision or positive predicted value (PPV) and recall or true predicted value (TPR). Those metrics are commonly

used to improve the process of model evaluation when dealing with imbalanced data. However, there is an inevitable trade-off between those two measures. As the ability of a model to find all GCU classes increases, the ability of a model to identify only GCUs decreases. We can notice that TPRs are relatively small for all models, never greater than 50%, and extremely low for plugin lasso (3.8%). This also means that false negative rates (FNRs) or type II errors are high (FNR=1-Recall). While traditional *LRfull* (47.2 %) and *adaptive* lasso (45.3%) have the highest recall, *ridge regression* has the greatest precision (83.7%).

However, when looking at selected imbalance-adjusted metrics (F-measure, Balanced accuracy, Geometric mean and Matthew's Correlation Coefficient), adaptive lasso performs slightly better than *LRfull*. Therefore, based on calculated classification metrics, regularization models have no clear advantage over traditional LR models.

Beside classification metrics based on nominal class prediction using a given threshold (in our case, standard 50%) threshold), we use two probabilistic classification metrics (ROC and precision-recall curve) that plot values for all possible thresholds. ROC analysis and AUC ROC are most commonly used to evaluate predictive performance [10]. ROC visualizes a nonlinear trade-off between TPR (recall or sensitivity) and FPR (1-specificity) values, and AUC summarizes this information into a single number, which facilitates model comparison when there is not a dominating ROC [30]. Table 4 shows the results of the ROC/AUC analysis for the test sample.

Model	**AUC**	**SE**	**95%-CI Lower**	**95%-CI Upper**
LRSW	0.8370	0.0317	0.7749	0.8991
LRFULL	0.8378	0.0314	0.7763	0.8994
lasso	0.8337	0.0314	0.7722	0.8952
adaptive	**0.8422**	0.0309	0.7816	0.9028
plugin	0.8166	0.0326	0.7527	0.8804
elastic net	0.8349	0.0313	0.7735	0.8963
ridge	0.8398	0.0326	0.7760	0.9037
Ha: At least one classifier has a different AUC value $\chi^2 = 4.60$, $p = 0.5961$				

Table 4: Area under the ROC

Notes: *The highest values of AUC ROC are in bold.*

The results again favor the *adaptive* lasso. *Adaptive* lasso has the highest AUC value (84%), while *plugin* lasso, again, has the lowest value (82%). However, the difference between the AUC values of the models in the test sample is not statistically significant (Table 4).

While AUC ROC is a popular metric, it can be inflated in the presence of class skew. Therefore, we compare the models' area under the precision-recall curve. In addition, this metric may be preferable if we are more concerned about the number of false negatives [10], which is the case for GCU prediction. Table 5 shows the values of precision-recall (PR) AUC. The results show that PR AUC values in test samples for different classifiers and under different sample split strategies (see Appendix B, available at: Appendix.pdf) are approximately the same. Again, there is no straightforward evidence that simpler or penalized models perform noticeably better than traditional unpenalized and overparametrized models.

Finally, we use calibration belts to examine the relationship between test sample estimated probabilities and observed GCU rates. Calibration gives insight into model uncertainty by adjusting the probability distribution to better match the expected distribution observed in the

Model	PR AUC
LRsw	0.5988
LRFULL	0.5805
lasso	0.5771
adaptive	0.5932
plugin	0.5623
elastic net	0.5789
ridge	**0.6014**

Table 5: Area under the Precision-recall curve

Notes: *The highest values of AUC ROC are in bold.*

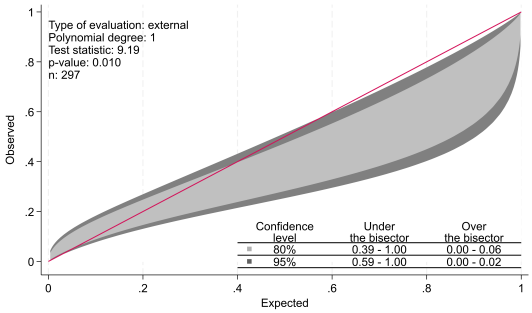

Figure 1: LRsw

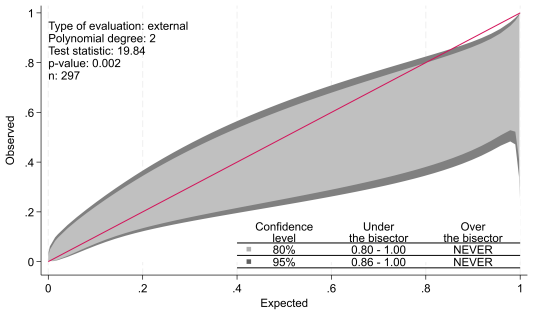

Figure 2: LRfull

data. Calibrated predictions imply that among those observations assigned a predicted rate of, for example, 0.10 for a GCU, the actual GCU does occur at a 10% rate. We present calibration belts using "calibrationbelt" command in Stata, which implements the calibration belt and its associated test (see: Nattino et al., 2017 [22]). Figures from 1 to 7 show calibration belts for estimated models on the test sample and corresponding test statistics of the deviations from the line of perfect calibration.

As can be noticed from Figures 1-7, only a few models (*lasso, elastic net and ridge regression*) showed satisfactory external calibration in the test sample. Because they are well-calibrated, the predictions of those models have greater economic significance.

Overall, our results indicate that, without considering calibration results, we would not see a clear cut between the traditional and sophisticated regularization models' performance and might choose an unstable predictive model. For example, while classification metrics favor adaptive lasso and LRfull GCU models, even after taking into account imbalance-adjusted

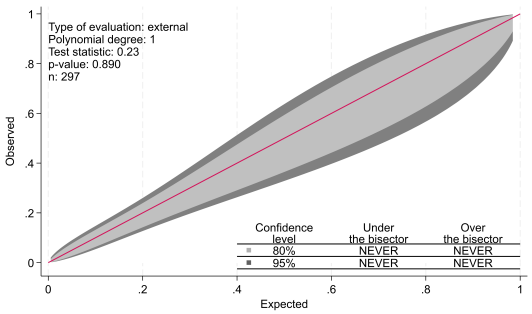

Figure 3: LASSO CV

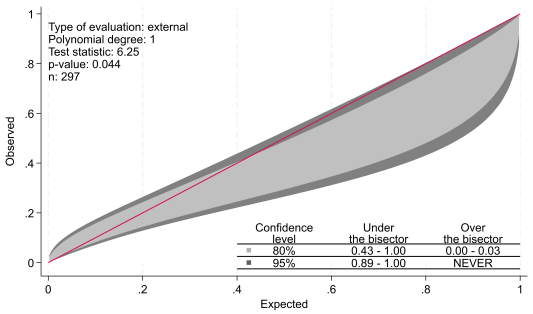

Figure 4: ADAPTIVE LASSO

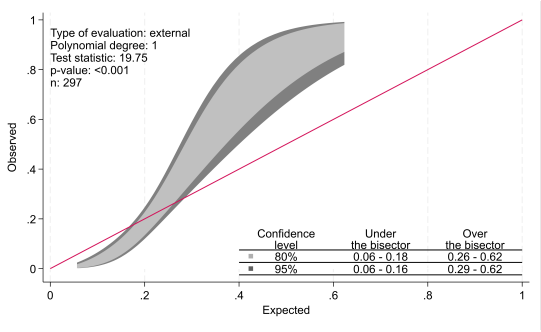

Figure 5: PLUGIN LASSO

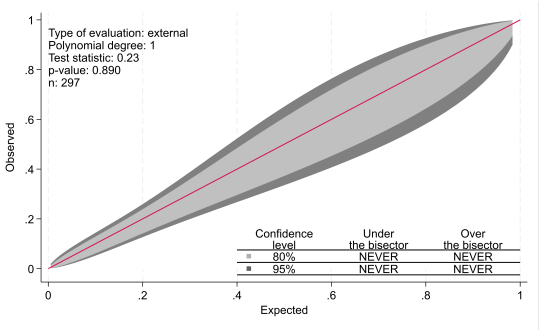

Figure 6: ELASTIC NET

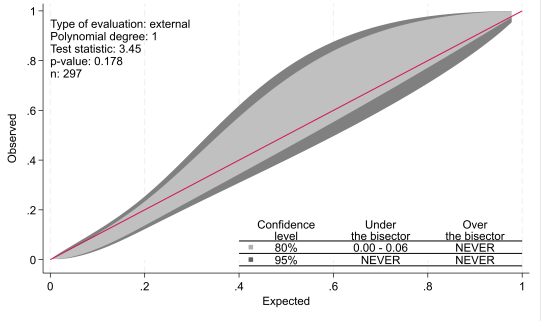

Figure 7: RIDGE REGRESSION

classification metrics, they may significantly underestimate small risks and/or overestimate high risks. In this regard, calibration aims to prevent predicted risks from being too extreme and makes the predictive model more relevant. Finally, we used an alternative sample split (80:20%) to train the models with more data. Our results are robust to alternative sample split decisions (see Appendix B, available at:Appendix.pdf).

5. Summary and Conclusion

Although research on different default prediction models has been around for quite a long time, new technologies and big data have opened the door for the implementation of more advanced and powerful prediction models. However, choosing the proper risk prediction model based on imbalanced datasets is rather challenging, even for sophisticated ML methods. Such datasets can cause traditional, as well as sophisticated classification algorithms to have a biased decision boundary. Our results show that the traditional, unregularized, and likely overfitted models (*LRsw* and *LRfull*) may perform as well as sophisticated regularization models (*lasso, adaptive lasso, plugin lasso, elastic net and ridge regression*) in terms of their classification properties when facing imbalanced datasets. However, we find that the models with the highest performance accuracy (*adaptive* and *LRfull*) are not well-calibrated by default, therefore lacking sufficient confidence in their GCU predictions. This implies that predictive model evaluation should be carried out carefully, preferably over the range of classification and calibration metrics and, if possible, taking into account the costs of prediction errors. While previous studies on GCU prediction have investigated a wide variety of financial and non-financial predictors using different approaches to variable selection, different sample splitting and selection methodologies, different model evaluation criteria, little or no attention, to the authors' best knowledge, has been given on assessing models' uncertainty. Our results show that the final decision in model

selection should consider the model calibration results and a combination of different classification performance metrics. Further efforts could be made to explore potential heterogeneity and more profoundly investigate the generalizability of model performance. For example, instead of focusing on undersampling, oversampling or different sample matching techniques, researchers using sample sizes that are large enough (which is our limitation) could instead focus on potentially interesting subdomains (e.g. companies with indications of poor financial performance) without trying to remove imbalance via sampling methods artificially.

Acknowledgements

This work has been fully supported by Croatian Science Foundation under the project IP-2020-02-9372 "Disentangling Financial Reporting Quality".

References

[1] Ahrens, A., Hansen, C. B. and Schaffer, M. E. (2020). lassopack: Model selection and prediction with regularized regression in Stata. The Stata Journal, 20 (1), 176 - 235. doi: 10.1177/1536867X20909697

[2] Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. Annual Review of Economics, 11, 685 - 725. doi: 10.1146/annurev-economics-080217-053433

[3] Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica, 80(6), 2369 - 2429. doi: 10.48550/arXiv.1010.4345

[4] Belloni, A., Chernozhukov, V. and Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. Journal of Business and Economic Statistics, 34 (4), 606 - 619. doi: 10.1080/07350015.2016.1166116

[5] Bellovary, J. L., Giacomino, D. E. and Akers, M. D. (2007). A review of bankruptcy prediction studies: 1930 to present. Journal of Financial Education, 1 - 42.

[6] Bertomeu, J. (2020). Machine learning improves accounting: discussion, implementation and research opportunities. Review of Accounting Studies, 25(3), 1135 - 1155. doi: 10.1007/s11142-020-09554-9

[7] Blagus, R. and Goeman, J. J. (2018). What (not) to expect when classifying rare events. Briefings in Bioinformatics, 19 (2),. 341 - 349. doi: 10.1093/bib/bbw107

[8] Carson, E., Fargher, N. L., Geiger, M. A., Lennox, C. S., Raghunandan, K. and Willekens, M. (2013). Audit reporting for going-concern uncertainty: A research synthesis. Auditing: A Journal of Practice and Theory, 32(1), 353 - 384. doi: 10.2308/ajpt-50324

[9] Chye Koh, H. and Kee Low, C. (2004). Going concern prediction using data mining techniques. Managerial Auditing Journal, 19 (3), 462 - 476. doi: 10.1108/02686900410524436

[10] Cook, J. and Ramadas, V. (2020). When to consult precision-recall curves, The Stata Journal, 20 (1), 131 - 148. doi: 10.1177/1536867X2090969

[11] Curth, A., Jeffares, A. and van der Schaar, M. (2024). A U-turn on double descent: Rethinking parameter counting in statistical learning. Advances in Neural Information Processing Systems, 36. doi: 110.48550/arXiv.2310.18988

[12] Čorić, R., Matijević, D. and Marković, D. (2023). PollenNet-a deep learning approach to predicting airborne pollen concentrations. Croatian Operational Research Review, 14(1), 1-13. doi: 10.17535/crorr.2023.0001

[13] Del Giudice, M. (2024). The prediction-explanation fallacy: a pervasive problem in scientific applications of machine learning, Methodology, 20(1), 22 - 46. doi: 10.5964/meth.11235

[14] Gadžo, A., Suljić, M., Jusufović, A., Filipović, S. and Suljić, E. (2025). Data mining approach in detecting inaccurate financial statements in government-owned enterprises. Croatian Operational Research Review, 16(1), 1-15. doi: 10.17535/crorr.2025.0001

[15] Goo, Y. J. J., Chi, D. J. and Shen, Z. D. (2016). Improving the prediction of going concern of Taiwanese listed companies using a hybrid of LASSO with data mining techniques. SpringerPlus, 5(1), 1 - 18. doi: 10.1186/s40064-016-2186-5

[16] Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction, Springer, New York.
[17] Hastie, T., Tibshirani, R., Wainwright, M. (2015). Statistical learning with sparsity, Monographs on statistics and applied probability 143, CRC Press, Boca Raton.
[18] Hsu, Y. F. and Lee, W. P. (2020). Evaluation of the going-concern status for companies: An ensemble framework-based model. Journal of Forecasting, 39(4), 687-706. doi: org/10.1002/for.2653
[19] Kalinić Milićević, T., and Marasović, B. (2023). What factors influence Bitcoin's daily price direction from the perspective of machine learning classifiers? Croatian Operational Research Review, 14(2), 163-177. doi: 10.17535/crorr.2023.0014
[20] Krupa, J. and Minutti-Meza, M. (2022). Regression and Machine Learning Methods to Predict Discrete Outcomes in Accounting Research. Journal of Financial Reporting, 7(2), 131-178. doi: 10.2308/JFR-2021-010
[21] Martens, D., Bruynseels, L., Baesens, B., Willekens, M. and Vanthienen, J. (2008). Predicting going concern opinion with data mining. Decision Support Systems, 45(4), 765-777. doi: 10.1016/j.dss.2008.01.0
[22] Nattino, G., Lemeshow, S., Phillips, G., Finazzi, S. and Bertolini, G. (2017). Assessing the Calibration of Dichotomous Outcome Models with the Calibration Belt, The Stata Journal, 17(4), 1003-1014. doi: 10.1177/1536867X1801700414
[23] Saeedi, A. (2021). Audit opinion prediction: A comparison of data mining techniques. Journal of Emerging Technologies in Accounting, 18(2), 125 - 147. doi: 10.2308/JETA-19-10-02-40
[24] Saeedi, A. (2023). A High-Dimensional Approach to Predicting Audit Opinions. Applied Economics, 55(33), 3807 - 3832. doi: 10.1177/1536867X1801700414
[25] Shmueli, G. (2010). To explain or to predict?, Statistical Science, 25(3), 289-310. doi: 10.1214/10-STS330
[26] Šinkovec, H., Heinze, G., Blagus, R. and Geroldinger, A. (2021). To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. BMC medical research methodology, 21, 1 - 15. doi: 10.1186/s12874-021-01374-y
[27] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267 - 288. doi: 10.1111/j.2517-6161.1996.tb02080.x
[28] Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L. and Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. BMC medicine, 17(1), 1- 7. doi: 10.1186/s12916-019-1466-7
[29] Weiss, G. M. (2013). Foundations of imbalanced learning. In: He, H., and Ma, Y. (Eds.). Imbalanced learning: Foundations, Algorithms, and Applications, Hoboken, NJ, USA: Wiley, 13 - 41. doi: 10.1002/9781118646106.ch2
[30] Yang, Z., Yu, Y., You, C., Steinhardt, J. and Ma, Y. (2020). Rethinking bias-variance trade-off for generalization of neural networks. In H. Daumé, III and A. Singh (Eds.), Proceedings of the 7th International Conference on Machine Learning, pp.10767–10777. url: https://proceedings.mlr.press/v119/yang20j.htm
[31] Yeh, C. C., Chi, D. J. and Lin, Y. R. (2014). Going-concern prediction using hybrid random forests and rough set approach. Information Sciences, 254, 98 - 110. doi: 10.1007/978-3-319-09333-8_24
[32] Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101(476), 1418 - 1429. doi: 10.1198/016214506000000735