Center for
Economic Theories and Policies
Sofia University St. Kliment Ohridski
Faculty of Economics and Business Administration

ISSN: 2367-7082

Bulgarian Economic Papers

Binary Classification Problems in Economics and 136 Different Ways to Solve Them

Anton Gerunov

BEP 02-2020
Publication: March 2020

Binary Classification Problems in Economics and 136 Different Ways to Solve Them

Anton Gerunov[1]

Abstract

This article investigates the performance of 136 different classification algorithms for economic problems of binary choice. They are applied to model five different choice situations – consumer acceptance during a direct marketing campaign, predicting default on credit card debt, credit scoring, forecasting firm insolvency, and modeling online consumer purchases. Algorithms are trained to generate class predictions of a given binary target variable, which are then used to measure their forecast accuracy using the area under a ROC curve. Results show that algorithms of the Random Forest family consistently outperform alternative methods and may be thus suitable for modeling a wide range of discrete choice situations.

Key Words: discrete choice, classification, machine learning algorithms, modeling decisions

JEL: C35, C44, C45, D81

[1] Associate Professor at Faculty of Economics and Business Administration, Sofia University "St. Kliment Ohridksi", e-mail: a.gerunov@feb.uni-sofia.bg

I. Introduction

Problems of binary choice and classification are prevalent in all fields of economics and business. Predicting consumer choice, credit risk scoring, investigating online and offline behavior, and business decision-making are merely a few examples in the long list of possible applications (Hensher & Johnson, 2018). Early attempts to statistically model binary choice have given us instruments like the logistic regression but recent advances in statistics and machine learning have produced a large and diverse set of classification methods. While their uptake as standard econometric instruments has not kept pace with the rapid speed of advancements in machine learning, their potential utility is already being demonstrated in recent research (e.g. Gerunov, 2019). However, the benefits of using such methods are partly offset by their large number, complexity and limited application to economic problems. In short, there is insufficient research on which of those novel methods can be fruitfully applied to economic problems and how their classification accuracy stacks against other alternatives.

This article sets out to close this gap by investigating 136 different machine learning methods by applying them to five common economic problems. Those problems range from predicting consumer choice during direct marketing, through credit risk scoring, estimating credit card and company default rates, to predicting online purchasing behavior. We probe what families of algorithms produce highest accuracy and give insight into their application to the fields of economics and business. The article is structured as follows: the second section provides an overview of relevant literature, the third section defines criteria for classification accuracy, the fourth section briefly presents the methodology and data used. The fifth section contains the results, while the sixth discusses and compares them. The seventh section concludes.

II. Literature Review

Using rigorous statistical methods for modeling binary choice can be traced back at least to Cox's (1958) pioneering work and probably even before that. Cox (1958) introduced the logistic regression, which was refined over the next decades (Manski and McFadden, 1981) and even today is a preferred tool for classification in relatively well-understood problems using small to medium-

sized samples (Hyman and Yang, 2001; Akinci et al., 2007; Jie et al., 2019). The premise of the choice problem is clear. A given observations needs to be classified as belonging to one of two classes (e.g. positive and negative), using its known features.

In this sense, it is useful to have labeled data where observations are divided into two groups, the first of which constitutes the first class, while the second consists of the other one. In terms of the information structure of each observation, we emphasize that it is appropriate to store it together with the associated circumstances (variables, characteristics) and, if possible, indicate to which of the two groups it belongs. An example of this could be a series of financial transactions with their specific characteristics that are designated as legitimate or fraudulent. The utility of discrete classification methods is even larger than it seems since a lot of continuous problems can be collapsed into problems of discrete choice by grouping a continuous variable. For instance, in the case of analysis of continuous variables (eg temperature, hours of operation, costs), anomalous observations can be detected by defining a criterion and corresponding value over which we label them as different from the default class.

Later work introduced the linear discriminant analysis (LDA) as another approach to distinguish between classes in economic and business problems (Ripley & Hjort, 1996). This method has been extensively and fruitfully used to model situations of consumer choice (Tregear and Ness, 2005; Hansen, 2005). While both the logistic regression and the LDA has led to many insightful results, their utility has been put to the test in an era of rapidly expanding data availability and more complex research problems. Most notably, classical methods are intrinsically linear in nature, while complex economic phenomena often tend to exhibit non-linear relationships. Second, those methods stem from a rich statistical tradition but often impose stringent assumption on the data structure. Finally, traditional methods often do not scale well to large or extremely large sample sizes (so-called “big data”). This is particularly notable in the estimation of p-values are they are biased towards significance in such a setting (Greenland, 2019).

In partial remedy of those concerns, one can peruse novel classification methods from the field of machine learning and suitably apply them to forecast group association (Zhao et al., 2014). Since this classification problem calls for labeled data, the group of supervised methods are used most

often. Supervised machine learning algorithms are characterized in that they need labeled data with clearly delineated classes (or values) of the target variable. This most often involves human or machine data processing to determine whether the target variable belongs to a normal class (negative) or belongs to an anomalous class (positive). Although the simplest markup is a binary (dummy) variable with class 0 (normal, legitimate), 1 (risk, fraud, problem, deviation), there is no reason why labels do not have more meaning to account for the nuances of conversions. For example, Gerunov (2016) looks at modeling large data sets to evaluate the risk of unemployment. Individuals may be marked as employed (class 0) or unemployed (class 1), but in order to achieve greater detail and a clearer understanding of the processes, the two classes are divided into different subclasses (employed, self-employed, partially employed, unemployed, pensioners, etc.) A random forest classification model is trained on this data, and it is able to successfully identify the drivers of unemployment.

Machine learning algorithms are used for a wide range of different economic and business problems. Algorithms such as the Naïve Bayes classification have been used for modeling consumption choice and user sentiment (Ye et al., 2009; Cheung et al., 2003; Huang et al., 2012). The rise in application of the support vector machine (SVM) algorithms is also notable. SVMs have been applied to numerous tasks such as consumer preference elicitation (Huang & Luo, 2016) and sentiment analysis (Hariguna & Romadon, 2019). Decision trees and random forests have also been extensively applied to problems of economic choice. This application ranges from deriving consumer preferences (Bi, 2012) and modeling decisions (Kruppa et al., 2013) to credit risk management (Meng et al., 2019). Finally, the recent upsurge in interest in neural networks has also produced a large number of applications to economic problems – e.g. in predicting consumer decision-making (Reunolds & Philips, 2019), credit scoring (Fu et al. 2016), consumer loyalty (Deliana & Rum, 2017), and many others.

Currently, the most commonly used algorithms for supervised classification in the research literature and practice are neural networks, k-closest neighbors, base networks, trees and decision trees, support vector machines, but also traditional statistical approaches. such as logistic regression and discriminant analysis (Chandola et al., 2009; Phua et al., 2010; Omar et al., 2013; Qiu et al., 2016; Rousseeuw & Hubert, 2018). We emphasize that there are many new algorithms

that can be potentially useful, and also that many of them have variations of those already listed. Chandola et al. (2009) note that, as a general rule, supervised learning algorithms are more accurate than non-supervised learning algorithms.

Choosing the optimal classification algorithm in terms of classification accuracy is often a challenging task. On the one hand, it is important for the algorithm to have good predictive power by correctly classifying a significant portion of the observations. On the other hand, classification errors sometimes have different costs - for example, a borrower who is misclassified as unreliable leads to foregone profit, while one that is incorrectly classified as a trustworthy is a loss. Thus, it is important to pay attention not only to the overall accuracy of the classification, but also to more detailed indicators of the algorithm's qualities. Additionally, non-technical requirements such as comprehensibility, ease of interpretation, or compliance with established practices or regulatory requirements may also be imposed when selecting such a criterion.

All those considerations leave the academic and the practitioner with a large amount of potential algorithm choices for a given classification problem and little guidance on how to proceed. This has led some researchers such as Fernandez-Delgado et al. (2014) to ask whether we indeed need hundreds of algorithms to solve problems of choice. Fernandez-Delgado et al. (2014) investigate a large number of classification algorithms and ranks them according to their probabilities of achieving maximum accuracy (PAMA).

The authors (ibid.) show that Random Forest implementations and SVMs achieve highest PAMA, followed by neural networks and boosted ensembles. These results are enlightening and an excellent initial foundation for further work but it would be useful to see more detailed algorithm performance in concrete decision problems. This includes not merely accuracy but also resource-intensiveness of the method. Even more importantly, classification accuracy may need to be more carefully estimated by fully taking into account correctly classified and misclassified observations. To this end we propose to use the Receiver Operating Characteristic (ROC) curve that is described in more detail in the following section.

III. Defining Measures of classification accuracy

Problems of binary choice are often connected to high-stakes decisions with potentially large impact, which is why achieving high accuracy is of significant importance. The overall classification accuracy shows what proportion of the model's predictions are correct and what proportion is not (Mateev, 2016). Apart from overall classification accuracy, we are often interested in accuracy per each of the given classes. This is summarized the classification matrix, presented in Table 1 (Kabakchieva, 2012; Semerdjieva et al., 2013).

Table 1: Classification Matrix

		True Class	
Predicted class		1	0
	1	True positive, *TP*	False positive, *FP*
	0	False negative, *FN*	True negative, *TN*
	Total	Positive, *P*	Negative, *N*

Based on these ratios, we can define a series of indicators for the predictive accuracy of a classification model (Fawcet, 2004). First of all, we take into account the overall balanced accuracy:

$$BA = \frac{TP + TN}{P + N} \tag{1}$$

The precision *Pr* is an indicator that allows us to evaluate the ability of the classification algorithm to correctly identify the positive classes. Precision is defined as follows:

$$Pr = \frac{TP}{TP + FP} \tag{2}$$

Similarly, the sensitivity *Ss* allows us to estimate what percentage of all positive-class observations are correctly identified, or:

$$Ss = \frac{TP}{TP + FN} = \frac{TP}{P} \tag{3}$$

Sensitivity is sometimes referred to as the proportion of real positive observations. Specificity, *Sp*, indicates how well the algorithm can correctly identify negative observations by measuring what percentage of all negative observations are successfully predicted as such. Specificity is also known as the proportion of actual negative class observations. The definition of specificity is as follows:

$$Sp = \frac{TN}{FP + TN} = \frac{TN}{N} \tag{4}$$

The metrics shown so far show different measures of predictive accuracy, i.e. to what extent the model does well with the classification task. Alternatively, we can deduce a measure of mistakes. The total percentage of erroneously predicted observations (error), *E*, is defined as follows:

$$E = \frac{FP + FN}{P + N} \tag{5}$$

The F-measure is sometimes used to evaluate the predictive accuracy of a classification algorithm:

$$F = \frac{2}{\frac{1}{Pr} + \frac{1}{Ss}} = \frac{2}{\frac{TP + FP}{TP} + \frac{P}{TP}} = \frac{2TP}{(TP + FP + P)} \tag{6}$$

Among the quality indicators of a forecast model, it is worth mentioning the kappa statistics (Carletta, 1996). The availability of a wide range of different indicators for evaluating classification models implies some freedom for researchers to make the final call.

As an alternative measure of the quality of a classifier, we can use the area under the Receiver Operating Characteristic (ROC) curve (Walter, 2005). The ROC space is a two-dimensional space that shows how a classifier is represented by the proportion of true positive observations (sensitivity, *Ss*) and the proportion of false positive observations *FP / N*. The first is a measure of the benefit of a given classifier, and the second a measure of its cost. The point classification resulting from the application of an algorithm can be used to calculate the sensitivity and the proportion of misclassified observations. These two indicators set the coordinates of the point of the algorithm in the ROC space (see Figure 1).

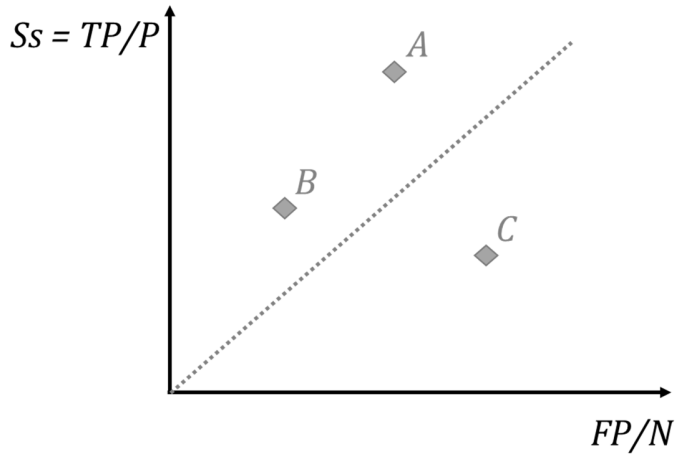

Figure 1: ROC Space with Three Alternative Classifiers

It is worth noting three main points in the ROC space:

- Origin point with coordinates (0; 0) – corresponds to cases where the algorithm never defines a class as positive;
- Point with coordinates (1; 1) – corresponds to cases where the algorithm always defines a class as negative;
- Point with coordinates (0; 1) – corresponds to the best possible classifier, which always determines the correct class.

The 45-degrees lines starting at (0; 0) and ending at (1; 1), corresponds to a completely random classification - algorithms that select a class solely because of chance fall into it. Therefore, any realistic and useful classifier should fall above this line, i.e. be better than classification by chance (e.g. algorithms A and B). Classifiers below the line perform worse than chance and should not be used (e.g. algorithm C). Many classification algorithms derive a probability distribution for class belonging or calculate some sort of similar test statistics. This allows the ROC space to show not just the point representation of the algorithm, but an entire curve reporting the results of the algorithm at different parameter values or test statistics. This curve is precisely the ROC curve (see Figure 2: Performance curve (ROC curve), and the area below it is a measure of how accurate the classification is (Fawcet, 2004).

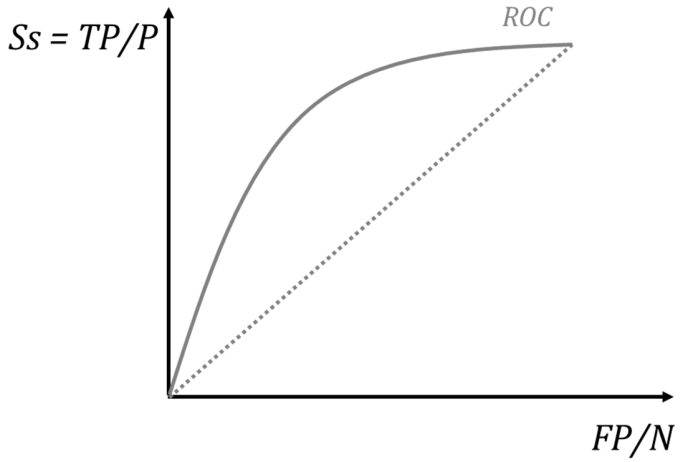

Figure 2: Receiver Operating Characteristic (ROC) Curve

The area under the curve (AUC) takes into account the tradeoff between the generated benefits of the classifier and the errors made, thus providing a single indicator that can be used for comparison between alternative classification models (Walter, 2005; Tanwani et al, 2009). The value of this area varies between 0 and 1. It should be borne in mind that the diagonal in space has $AUC = 0.5$, so we expect useful classifiers to have $AUC > 0.5$. Hand & Till (2001) also show the relationship between the AUC and another important indicator of the quality of the classification - Gini coefficient, G (see Breiman, 1984):

$$G + 1 = 2 * AUC \tag{7}$$

The area under the ROC (AUROC) curve is widely regarded as a prime candidate for a single comparison metric that can be applied to a wide variety of alternative classification algorithms (Fawcet, 2004; Tanwani et al, 2009; Semerdjieva et al., 2013). This article also takes this approach and considers the AUROC as the leading classification accuracy metric since is adequately balances the tradeoff between instances of right and wrong classification. It also gives a unified metric that can be applied not merely to alternative methods in a given classification situations but also to compare accuracy across types of problems.

IV. Datasets and Methods under Investigation

The considerable number of current studies in statistical methods and machine learning also imply the rapid development of methods, approaches and algorithms that are available for research purposes. For the sake of comprehensive review that will ensure robust identification of optimal algorithms, we use five distinct datasets (see Table 2) to evaluate and test 136 of the most popular methods. The full list of methods is included in an appendix.

Table 2: Datasets and Sources

#	Testing Data Set	Binary Classification: Target Variable	Observations, N	Data Source
1	Direct Marketing Campaign	Customer Accept or Reject Offer	41,181	Moro et al., 2014
2	Taiwan Credit Card Operations	Service Credit Card Debt or Not	30,000	Yeh & Lien, 2009
3	German Bank Credits	Default or Non-default on Credit	1,000	Eggermont et al., 2004; Hofmann, 1994
4	Estimating Default Status	Firm Default or Not	5,910	Zieba et al., 2016
5	Online Customer Purchases	Customer Purchase or Not	12,330	Sakar et al., 2018

More specifically we focus on the following types of problems:

- **Predicting customer acceptance during a direct marketing campaign** – the stems from a Portuguese bank and consists of known features of customers that can be used to glean insight whether customer will accept promotional offer.
- **Predicting default rates on credit cards** – the data from Taiwan given information about credit card accounts and customer and is focused on identifying which customer is likely to default on its debt.
- **Risk scoring for credits** – this task is focused on classifying which lenders are likely to service their loans and which are not. To this end we leverage Hofmann's (1994) data and use all the customer features to make the class prediction.

- **Estimating default status and probability** – this dataset gives detailed financial information for a large sample of Polish companies that can be used to classify them as likely to default or solvent.
- **Online Customer Purchases** – the dataset provides information for users of an online shop and their actions on a given website. These features can then be leveraged to identify which customers will purchase something, and which will not.

While this article investigates five main problems of binary choice but the conclusions obtained can easily be generalized and extended to other classification situations in economics, business, and possibly, beyond. The datasets, classification problems, observations, and original references are summarized in Table 2. The reader is directed to the original papers for further details on the data and its structure. The first three problems are also described and modeled in Gerunov (2020), where the author applies a total of seven algorithms and compares their performance.

Using the five datasets under investigation we select 136 classification algorithms and use all the features of the original data to train a model on them. In order to minimize the risk of overfitting and produce reliable results, we divide data into two subsets – one for training (the train set that consists of 80% of the sample), and one for testing (the test set that is the remaining 20% of the sample). We only use the test set to obtain out-of-sample classifications. In this testing, we calculate the area under the ROC curve of each algorithm, and also measure its complexity. As an approximation to the measure of complexity, we use the time required to calculate a given algorithm, standardizing the longest required time at 100% and presenting the remaining times as fraction of it. Therefore, the measure of complexity varies between 0% and 100%.

The computation time of an algorithm is highly dependent on the infrastructure used and the implementation method, and it is of particular importance whether the computation is distributed or not. It is misleading to report "raw" time as it will depend on the machines or clusters of machines used. The complexity measure partially solves this problem by resorting to relative numbers rather than absolute values. While there are some outstanding issues such as that computation time varies by processor type, architecture, load management, and other platform and infrastructure specific, the complexity measure is still a satisfactory approximation to how resource-intensive a given algorithm is.

V. Results

1. Direct Marketing Campaign Data

One of the main activities of a modern organization is attracting new customers or expanding relationships with existing ones. An example of the latter is conducting direct marketing initiatives to current customers, offering them a new product or service or an upgraded version of the one they are currently using. In these cases of direct marketing, the organization often has sufficient data before contacting customers so that their behavior can be modeled using statistical algorithms. The main problem with such campaigns is the realization of the risk of a client refusing the offered offers, so that the resources directed to that specific user are unproductive. While it is virtually impossible to achieve complete certainty as to whether a contact will lead to a successful sale, it is quite possible to model each potential contact and predict the probability of success. In this way, organizational resources can be directed to contacting customers with a high probability of success, thus minimizing the risk of unnecessary time and money spent. We thus use the algorithms under study to fit classification models with the offer acceptance as a target dependent variable.

The predictive accuracy of the methods considered is summarized in the histogram in Figure 3. First, the significant variation in accuracy between the different algorithms used is striking. Even if we ignore the extreme values (positive and negative), the bulk of the distribution changes from 0.5 to 0.78. This suggests that some algorithms are significantly better suited to a particular type of task than others. In this sense, choosing the right classification algorithm can lead to a very significant difference in the results generated and hence in the business value created. Secondly, we note that the distribution shown is close to normal, with a peak around AUROC = 0.7. This would be the expected predictive accuracy of the "average" algorithm that accomplishes this task. Third, we take into account the relatively high number of algorithms that do not add value (area below the curve of 0.5), emphasizing that they should be avoided.

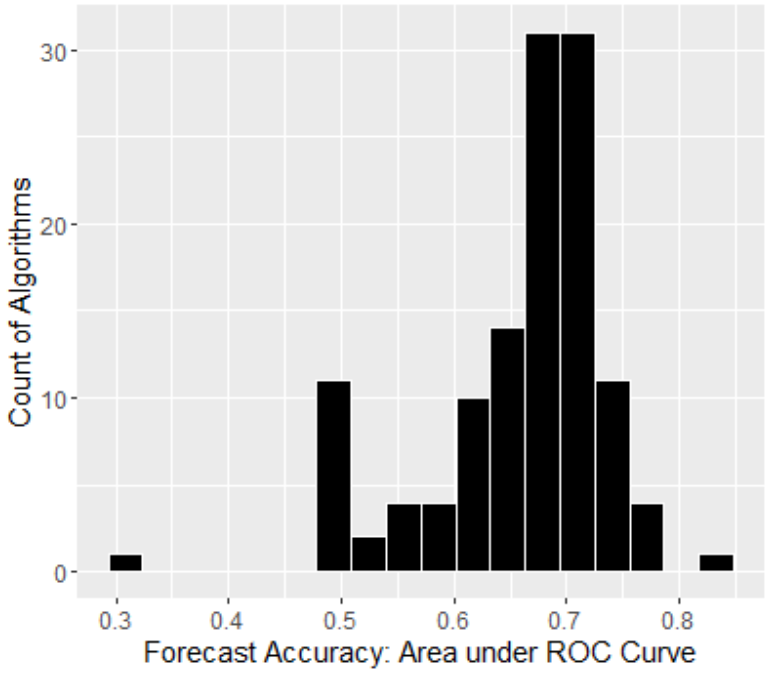

Figure 3: Histogram of Classifier Accuracy in Portuguese Marketing Campaign Data

The top ten classification methods with the best results are presented in Table 3. It is noteworthy that various models of discriminant analysis perform extremely well in this task, with the robust regularized linear discriminant analysis (rrlda) having an area under the operating characteristic curve of 0.82.

Table 3: Top 10 Most Accurate Classifiers for Direct Marketing Campaign Data

Algorithm Type	**Implementation Method**	**Area under ROC Curve**	**Complexity Measure**
Robust Regularized Linear Discriminant Analysis	rrlda	0.824	1.8%
Soft Independent Modeling of Class Analogies, SIMCA	CSimca	0.773	0.3%
Rule-Based Classifier	JRip	0.767	7.8%
Mixture Discriminant Analysis	mda	0.764	0.1%
Conditional Inference Tree	ctree	0.762	0.3%
C4.5-like Trees	J48	0.754	3.0%
Model Averaged Neural Network	avNNet	0.740	3.9%
ROC-Based Classifier	rocc	0.738	0.7%
Bagged AdaBoost	AdaBag	0.734	2.3%
Tree-Based Ensembles	nodeHarvest	0.734	56.7%

Decision trees and different types of ensemble algorithms also rank among the top ten classifiers, with the area under the ROC curve for algorithms in places 2 to 10 varying from 0.73 to 0.77. We also emphasize that the most accurate classification is not achieved by the most resource-intensive algorithms, and the top 10 include optimized and relatively fast algorithms.

2. *Credit Card Debt Data*

Credit card debt service is a key problem in the financial sector, as failure to do so and potential fraud can have a significant effect on the financial flows and solvency of their dependent organizations. In this context, it is particularly important to choose the optimal algorithm, and even small improvements in predictive accuracy can lead to unlocking significant value for lenders. For this purpose, we perform a comprehensive testing of 136 basic algorithms in the field of machine learning and analyze their accuracy in classification.

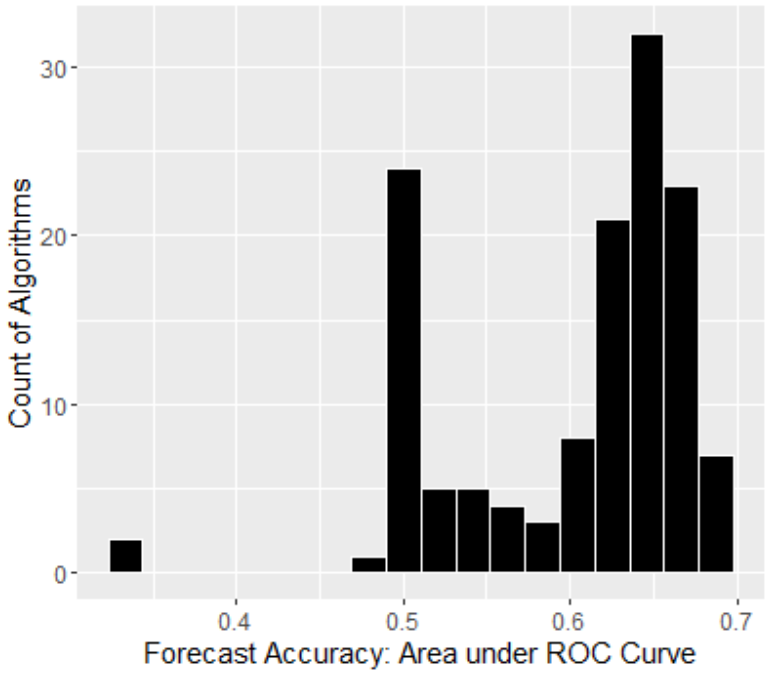

Figure 4: Histogram of Classifier Accuracy in the Taiwan Credit Card Debt Data

Figure 4 summarizes the predictive accuracy data of the alternative algorithms, measured as the area under the performance curve (ROC curve). A large number of the considered algorithms have an AUROC of about 0.50, which is a result equal to the chance - therefore this first peak of the

distribution shows the unfavorable algorithms for this task. In the histogram, we also observe a second peak with values around 0.65, with the vast majority of the algorithms considered concentrated precisely in the range of 0.62 to 0.68. The best algorithms tend to reach an area of 0.7, but in reality none exceeds this limit. We regard the classification task on credit card data as relatively difficult, which also explains the results obtained.

The top ten classification methods are presented in Table 2. It is noteworthy that the group is dominated by two main types of models – that of support vector machines and classification and regression trees (CART). The highest score is the polynomial kernel support vector machine, calculated using the least squares method with AUROC = 0.684, followed by the C5.0 type decision trees and three other variations of the support vector machines (all with AUROC - 0.682). The top ten list two more tree-based methods, one ROC-based classifier and an ensemble adaptive gain model.

Table 4: Top 10 Most Accurate Classifiers for Credit Card Debt Data

Algorithm Type	Implementation Method	Area under ROC Curve	Complexity Measure
Least Squares Support Vector Machine with Polynomial Kernel	svmPoly	0.684	1.1%
Single C5.0 Ruleset	C5.0Rules	0.682	0.1%
SVM Linear Weighted	svmLinearWeights	0.682	0.6%
SVM Linear	svmLinear	0.682	0.2%
SVM Linear2	svmLinear2	0.682	0.2%
ROC-Based Classifier	rocc	0.680	0.2%
CART	rpart1SE	0.679	0.1%
Bagged AdaBoost	AdaBag	0.677	37.4%
Boosted Tree	bstTree	0.677	1.9%
Boosted Classification Trees	ada	0.673	3.2%

We note that the accuracy of all these algorithms is very similar, and in practice relatively small differences will be observed, which would only be relevant when processing large data sets. As long as it takes time to calculate, the best algorithms are again not the most resource intensive. The optimal method is nearly one hundred times faster than the slowest one, the second best one is a thousand times faster and the third one is 167 times faster. This shows that in this task again we see an opportunity for balancing between the computational load and the accuracy of the results obtained.

3. German Credit Data

Credit risk modeling is a classic classification task and standard machine learning algorithms can be applied to it. We apply all the algorithms under study to this problem and present the summary results in Figure 5. The distribution observed in this case differs significantly from the normal one. We observe a peak of algorithms with predictive accuracy around chance (AUROC = 0.5), followed by a relatively uniform distribution of algorithms with predictive accuracy in the range 0.52 to 0.64. Many of the algorithms discussed have an accuracy in the range of 0.68 to 0.70, which can be said to be our expectation of an "average" algorithm suitable for this particular task. There are a small number of algorithms with AUROC> 0.70, which are the best performing classifiers for the target variable (loan repayment).

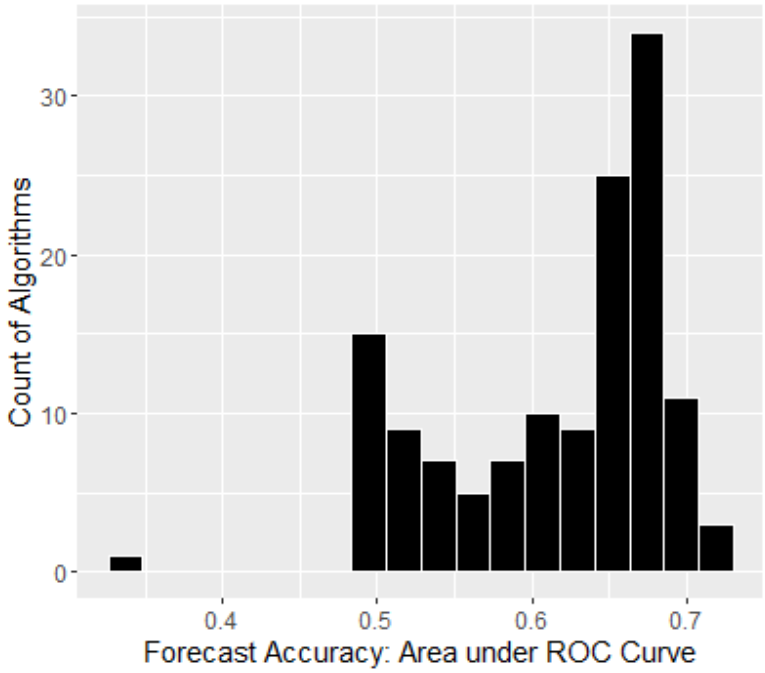

Figure 5: Histogram of Classifier Accuracy in the German Credit Data

The top ten classifiers with the highest predictive accuracy are presented in Table 4. All of them have an area below the ROC curve of at least 0.7, with the best being the regularized random forest (AUROC = 0.73). It is noteworthy that this group is dominated by implementation of the Random Forest algorithm that account for half of the top ten algorithms. Additionally, gradient boosting

methods, one specific type of neural network (multilayer perceptron), and a version of discriminant analysis (localized linear discriminant analysis) also perform very well. Again, the most computationally demanding algorithms do not produce the best results. The most accurate classifier is 2.6 times faster than the most resource-intensive one, and the second most accurate is 13.5 times faster.

Table 5: Top 10 Most Accurate Classifiers for German Credit Data

Algorithm Type	Implementation Method	Area under ROC Curve	Complexity Measure
Regularized Random Forest	RRF	0.730	38.5%
eXtreme Gradient Boosting	xgbLinear	0.718	7.4%
Regularized Random Forest	RRFglobal	0.712	5.2%
eXtreme Gradient Boosting	xgbDART	0.707	26.3%
Multi-Step Adaptive MCP-Net	msaenet	0.702	5.1%
Random Ferns	rFerns	0.701	2.0%
Localized Linear Discriminant Analysis	loclda	0.701	0.8%
Random Forest	rf	0.700	1.7%
Random Forest	ranger	0.698	1.6%
Gradient Boosting Machine	gbm	0.698	0.2%

4. Polish Companies Default Data

Determining whether a given company will default or not is a significant challenge. Using rich financial data on Polish companies, we are able to investigate the performance of different approaches to solving this classification problem. All the investigated models are calculated on a training sample and their predictions are tested on a test sample. The distribution of their accuracy, measured by the area under the performance curve, is summarized in the histogram Figure 6. A huge number (over 40) of the estimated algorithms have a predictive accuracy of about AUROC = 0.5, which is exactly equal to classification by chance. The difficulty of the task is also emphasized by the fact that there are some algorithms with an area under the ROC curve of less than 0.5, which is worse than a randomly generated forecast. We observe a slight peak in predictive accuracy at AUROC values of 0.6, with the best classification algorithms reaching AUROC predictive accuracy above 0.7. We note significant differences between the results of the different methods with only a minority of approaches displaying fairly high predictive accuracy.

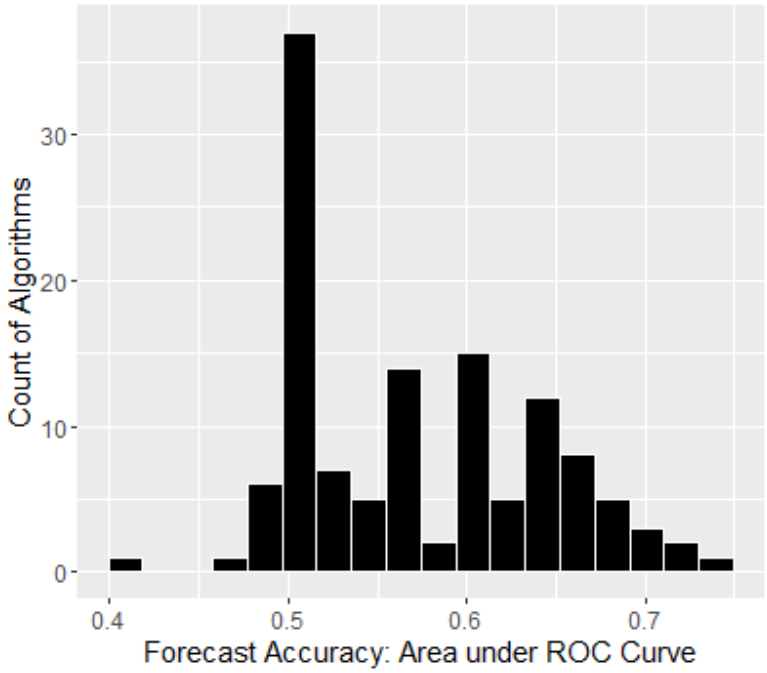

Figure 6: Histogram of Classifier Accuracy for Polish Company Defaults Data

The ten most accurate algorithms are presented in Table 5. The best performer among them is the robust soft independent modeling of class analogies, RSIMCA. This method is relatively obscure in the field of economics and business, but essentially involves supervised analysis that separates data into major components and constructs subspaces based on those components that are subsequently used for classification. For more details, we direct the reader to the original development of Brandon & Hubert (2005), as well as to the study of Fauziyah et al. (2018).

Table 6: Top 10 Most Accurate Classifiers for Polish Companies Default Data

Algorithm Type	Implementation Method	Area under ROC Curve	Complexity Measure
Robust SIMCA	RSimca	0.733	0.3%
Patient Rule Induction Method	PRIM	0.721	100.0%
Random Ferns	rFerns	0.714	1.7%
CART	rpart1SE	0.693	0.1%
CART	rpart2	0.693	0.0%
Single C5.0 Ruleset	C5.0Rules	0.693	0.2%
Rule-Based Classifier	PART	0.679	0.3%
Regularized Random Forest	RRF	0.676	35.6%
Shrinkage Discriminant Analysis	sda	0.676	0.2%
Bagged AdaBoost	AdaBag	0.674	0.9%

The RSIMCA model has an area under the ROC curve of 0.733 and is more than three hundred times faster than the slowest algorithm - the Patient Rule Induction Method, which is second in predictive accuracy with 0.721. Third, with very close predictive accuracy (AUROC - 0.714), ranks random trees, followed by six other methods in the decision tree or random forest family. They are all relatively fast and require relatively less computational resources. Tenth place is held by a specific method for discriminant analysis, which registers a relatively high accuracy – AUROC = 0.676.

5. *Online Purchases Data*

The final task is to find the optimal classification algorithm for online user behavior research. our goal here is to predict whether a customer in e-commerce setting will make a purchase or not. The summary results for algorithm predictive accuracy are shown in the histogram in Figure 7.

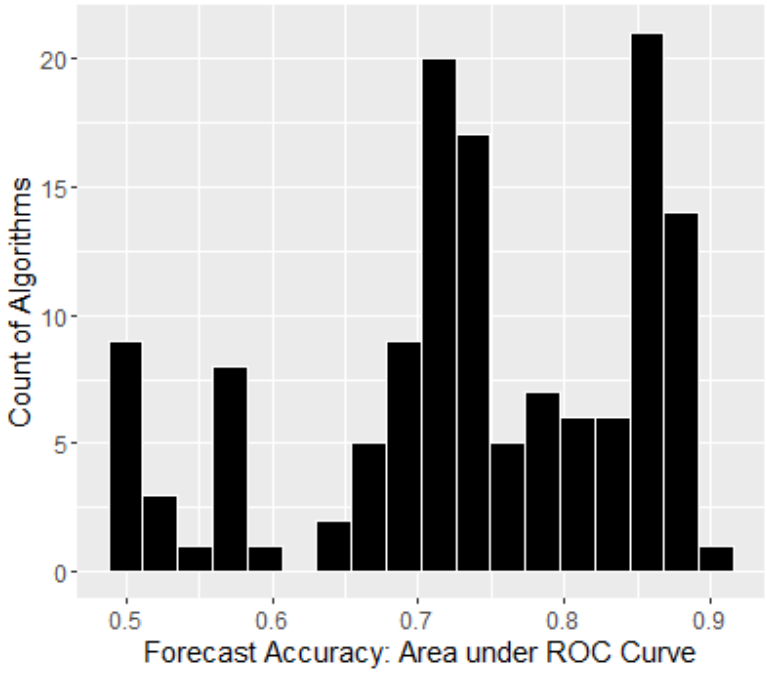

Figure 7: Histogram of Classifier Accuracy for Online Purchases Data

The average predictive accuracy in the classification of online behavior is significantly higher than the other situations considered. It is noteworthy that the AUROC distribution is characterized by

two peaks – one around AUROC = 0.70-0.75 and the other – around 0.85-0.90. The best classification algorithm scores even above 0.90. We also consider the significant variance in the results of the calculated methods. Very few of them have results close to chance, and a significant minority also report very high predictive accuracy. The ten most accurate algorithms are presented in Table 6. The rotation forest has the best performance with AUROC = 0.902, followed by the random forest with weighted spaces (0.889), the PRIM method (0.886), a boosted decision tree (0.885), and a series of methods from the classification and regression family (all with 0.880). We consider the family of classification and regression trees (CART), as the most optimal approach for solving this particular problem.

Table 7: Top 10 Most Accurate Classifiers for Online Purchases Data

Тип алгоритъм	**Метод**	**Площ под ROC-крива**	**Мярка за комплексност**
Rotation Forest	rotationForestCp	0.902	1.0%
Weighted Subspace Random Forest	wsrf	0.889	4.4%
Patient Rule Induction Method	PRIM	0.886	14.3%
Boosted Tree	bstTree	0.885	1.4%
CART	rpart	0.880	0.0%
CART or Ordinal Responses	rpartScore	0.880	1.4%
Conditional Inference Tree	ctree2	0.880	0.1%
C5.0	C5.0	0.879	0.3%
Cost-Sensitive C5.0	C5.0Cost	0.879	0.7%
DeepBoost	deepboost	0.876	4.9%

In terms of the time and resources required to calculate these methods, we emphasize that again the most resource-intensive methods do not produce the most accurate forecast. On the contrary, the optimal algorithm needs 100 times less time for computation than the slowest algorithm, and we observe similar and better ratios in the other methods as well. This underlines the possibility to simultaneously optimize both the accuracy and the IT resources and computing infrastructure required.

VI. Comparison and Discussion

The results obtained allows us to make a few preliminary observations on the applicability and utility of a wide range of machine learning classification methods. First, the best performing

algorithms are not the same for each of the considered problems – in each of the individual classification tasks a different approach performs best. This is probably due to the fact that different families of algorithms and their specific implementations are better suited to certain types of data, but they do worse with other types. This is a clear manifestation of the well-known no free lunch theorem (Branden & Hubert, 2005) and emphasizes that it is suboptimal to use one and the same method for every type of classification problem.

Second, we note that the predictive accuracy between different algorithms can vary significantly. Figure 8 summarizes the distribution of the averaged values of the area under the performance curve for all algorithms considered. Average predictive accuracy ranges from AUROC = 0.50 to 0.74. This emphasizes that the importance of choosing the optimal algorithm is not only theoretically justified but can have significant practical implications. This result further emphasizes the importance of comprehensively seeking out the best performing algorithms for a given task, as improvements in predictive accuracy have the potential to generate enormous business value given sufficient scale of application.

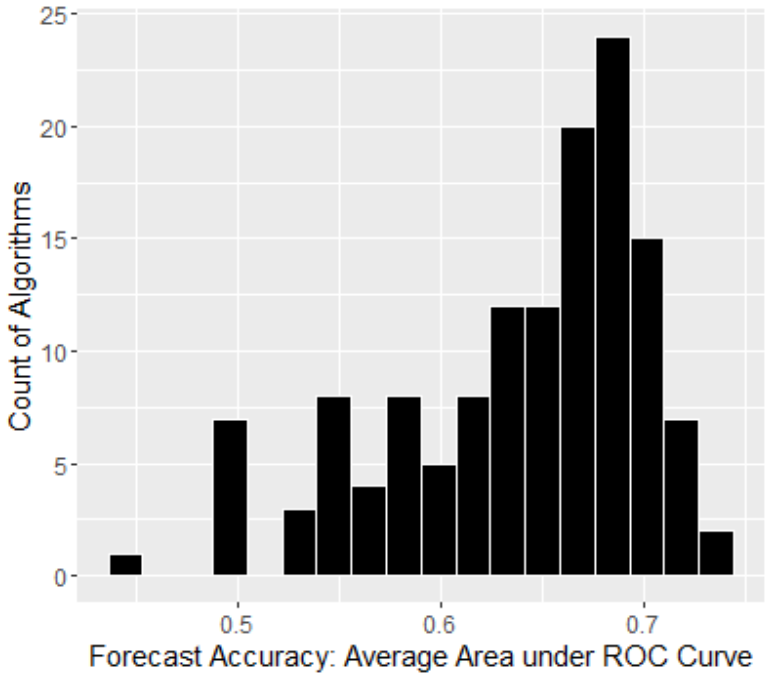

Figure 8: Histogram of Average Classifier Accuracy across All Problems

Third, we note that certain families of algorithms tend to perform better than others. In particular, different implementations of the Random Forest (or CART) family are often among the best peformers in solving each of the problems considered. In the analysis we noticed a common tendency for them to adjust to the particularities of the training data (i.e. to overfit it), but they still show excellent results in the test sample. Focusing on the traditional econometric tools, it seems that the linear discriminant analysis in its various implementations displays relatively good classification accuracy. Since this family of methods is in most cases highly optimized, they could be a reasonable compromise in situations where a considerable amount of data needs to be analyzed with limited computing resources.

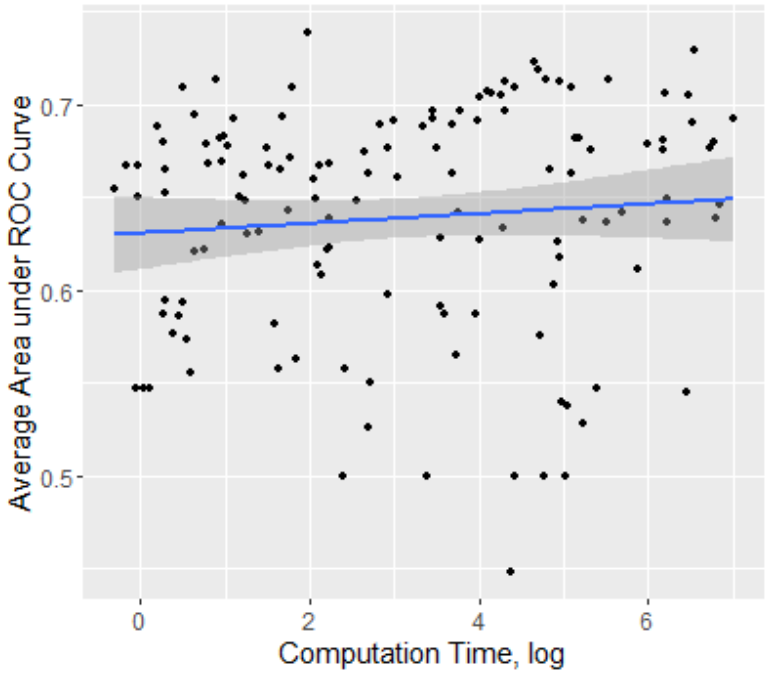

Figure 9: Relationship between Classification Accuracy and Estimation Time

Fourth, the most computationally intensive algorithms do not necessarily produce the most accurate class predictions. In each of the tasks considered, the best classification method is not the one that takes the most resources to calculate it. Figure 9 graphically presents this relationship between the mean area under the ROC curve for all methods on all tasks and the log of time required to evaluate them. The visual inspection reveals a weak positive relationship between the two, but when investigated within a linear regression model, this relationship does not reach

statistical significance ($p = 0.323$). Thus, it is possible to identify the optimal tradeoff between predictive accuracy and required computational resources in in order to generate as much value as possible from solving the classification task.

Results clearly show the potential for introducing novel machine learning algorithms to solve salient problems of discrete choice in the realm of business and economics. The research shows that we are able to identify novel methods that consistently outperform traditional econometric tools for binary classification such as the logistic regression or the linear discriminant analysis. Sometimes the differences in performance are large in size and such an improvement may have a potentially significant effect for practical applications. This holds particularly true for high-stakes decision situations such as credit risk scoring. Across all the reviewed methods and situations, we observed a robust trend that the family of Random Forest algorithms show consistently high performance, and thus recommend their more complete inclusion into the toolbox of standard econometric tools. More exotic methods sometimes do achieve somewhat higher classification accuracy, but this is often at the cost of increased computational resources. The linear discriminant analysis seems to be the algorithm of choice when solving big data classification problems with significant resource constraints.

VII. Conclusion

This short article tackles a basic question in modeling binary choices – what classification algorithm is likely to produce best results in terms of classification accuracy, thus enabling the researcher to glean more insight from data. Recognizing the wide variety of different and highly specific discrete choice problems in economics and business, we focused this exercise on five specific decision situations of assigning class: modeling consumer acceptance during direct marketing campaigns, predicting credit card debt defaults, credit risk scoring by a bank, predicting company defaults, and understanding online purchase decisions. Building upon data from previous research we tested a high number of alternative econometric and machine learning algorithms for classification and measured their performance.

This allows us to compare 136 of the most popular decision algorithms in terms of both their classification accuracy as measured by the area under the ROC curve, as well as in terms of their computational resource intensity. It seems that irrespective of the classification tasks, novel machine learning algorithms robustly outperform traditional econometric approaches such as the logistic regression. The latter more traditional methods are useful in situations of smaller datasets and limited computational resources, and even then the linear discriminant analysis should be preferred to the logistic regression. Among all the methods tested, implementations of the Random Forest (CART) family outperform almost any other method. This leads us to recommend their more thorough study as potential tool that can usefully complement the current econometrics toolbox.

REFERENCES

Akinci, S., Kaynak, E., Atilgan, E., & Aksoy, Ş. (2007). Where does the logistic regression analysis stand in marketing literature? A comparison of the market positioning of prominent marketing journals. *European Journal of Marketing*, *41*(5/6), 537-567.

Bi, J. (2012). A review of statistical methods for determination of relative importance of correlated predictors and identification of drivers of consumer liking. *Journal of Sensory Studies*, Vol. 27, No. 2, pp.87–101.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Belmont, CA: Wadsworth. *International Group*, *432*, 151-166.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, *22*(2), 249-254.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3), 15, 1-58.

Cheung, K.W., Kwok, J.T., Law, M.H. and Tsui, K.C. (2003). Mining customer product ratings for personalized marketing. *Decision Support Systems*, Vol. 35, No. 2, pp.231–243.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215-232.

Deliana, Y., & Rum, I. A. (2017). Understanding consumer loyalty using neural network. *Polish Journal of Management Studies*, *16*.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine learning*, *31*(1), 1-38.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *The Journal of Machine Learning Research*, *15*(1), 3133-3181.

Fu, K., Cheng, D., Tu, Y., & Zhang, L. (2016, October). Credit card fraud detection using convolutional neural networks. In *International Conference on Neural Information Processing* (pp. 483-490). Springer, Cham.

Gerunov, A. (2016). Employment modelling through classification and regression trees. *International Journal of Data Science*, *1*(4), 316-329.

Gerunov, A. (2019). Modeling Economic Choice under Radical Uncertainty: Machine Learning Approaches. *International Journal of Business Intelligence and Data Mining*, 14(1-2), 238-252.

Gerunov, A. (2020). Classification Algorithms for Modeling Economic Choice. *Economic Thought*. (forthcoming)

Greenland, S. (2019). Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *The American Statistician*, *73*(sup1), 106-114.

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning*, *45*(2), 171-186.

Hansen, T. (2005). Consumer adoption of online grocery buying: a discriminant analysis. *International Journal of Retail & Distribution Management*, Vol. 33, No. 2, pp.101–121.

Hariguna, T., & Romadon, Y. I. (2019, November). The accuracy comparison of vector support machine and decision tree methods in sentiment analysis. In *Journal of Physics: Conference Series* (Vol. 1367, No. 1, p. 012025). IOP Publishing.

Hensher, D. A., & Johnson, L. W. (2018). *Applied discrete-choice modelling*. Routledge.

Huang, B., Kechadi, M.T. and Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, Vol. 39, No. 1, pp.1414–1425.

Huang, D., & Luo, L. (2016). Consumer preference elicitation of complex products using fuzzy support vector machine active learning. *Marketing Science*, *35*(3), 445-464.

Hyman, M. R., & Yang, Z. (2001). International marketing serials: a retrospective. *International Marketing Review*, *18*(6), 667-718.

Jie, M. A., Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben van Calster. "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models." *Journal of clinical epidemiology* (2019).

Kabakchieva, D. (2012). *Investigating Data Mining Models for Classificaiton*. Ph.D, Thesis. Institute for Information and Communication Technologies, BAS. Sofia.

Kruppa, J., Schwarz, A., Arminger, G. and Ziegler, A. (2013). Consumer credit risk: individual probability estimates using machine learning. *Expert Systems with Applications*, Vol. 40, No. 13, pp.5125–5131.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of statistical software*, 28(5), 1-26.

Manski, C.F. and McFadden, D. (Eds.) (1981) *Structural Analysis of Discrete Data with Econometric Applications*, pp.2–50, MIT Press, Cambridge, MA.

Mateev, S. (2016). Evaluating methods for diagnostics and forecasting – analytic procedures and data interpretation. Sofia: New Bulgarian University.

Meng, C. Z., Liu, B. S., & Zhou, L. (2019, July). The Practice Study of Consumer Credit Risk Based on Random Forest. In *2019 International Conference on Modeling, Analysis, Simulation Technologies and Applications (MASTA 2019)*. Atlantis Press.

Omar, S., Ngadi, A., & Jebur, H. H. (2013). Machine learning techniques for anomaly detection: an overview. *International Journal of Computer Applications*, 79(2).

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint* arXiv:1009.6119.

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, *2016*(1), 67.

Reynolds, T. J., & Phillips, J. M. (2019). The Strata Model Predicting Advertising Effectiveness: A Neural-Network Approach Enhances Predictability of Consumer Decision Making. *Journal of Advertising Research*, *59*(3), 268-280.

Ripley, B. D., & Hjort, N. L. (1996). *Pattern recognition and neural networks*. Cambridge, UK: Cambridge university press.

Rousseeuw, P. J., & Hubert, M. (2018). Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(2), e1236.

Semerdjieva, V. Georgiev, B. & Damyanov, Ch. (2013). Data analysis of diagnostic tests. *Scientific Works of UFT*, 60, 292-297.

Tanwani, A. K., Afridi, J., Shafiq, M. Z., & Farooq, M. (2009). Guidelines to select machine learning scheme for classification of biomedical datasets. In *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*,128-139. Springer, Berlin, Heidelberg.

Tregear, A. and Ness, M. (2005). Discriminant analysis of consumer interest in buying locally produced foods. *Journal of Marketing Management*, Vol. 21, No. 1–2, pp.19–35.

Walter, S. D. (2005). The partial area under the summary ROC curve. *Statistics in medicine*, *24*(13), 2025-2040.

Ye, Q., Zhang, Z. and Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, Vol. 36, No. 3, pp.6527–6535.

Zhao, X., Shi, Y., Lee, J., Kim, H.K. and Lee, H. (2014).Customer churn prediction based on feature clustering and nonparallel support vector machine. *International Journal of Information Technology & Decision Making*, Vol. 13, No. 5, pp.1013–1027.

Appendix: Reference List of Algorithms Used

The statistical part of the paper is conducted in the R language for statistical computing, and the algorithm implementations are accessed through the R *Caret* package. For more information on the package we direct the reader to Kuhn (2008) and subsequent papers.

Table A1: A List of Tested Algorithms

#	Algorithm Name	Implementation Method
1	Adjacent Categories Probability Model for Ordinal Data	vglmAdjCat
2	Bagged CART	treebag
3	Bagged Flexible Discriminant Analysis	bagFDA
4	Bagged MARS	bagEarth
5	Bagged MARS using gCV Pruning	bagEarthGCV
6	Bayesian Generalized Linear Model	bayesglm
7	Boosted Generalized Linear Model	glmboost
8	Boosted Tree	blackboost
9	C5.0	C5.0
10	CART 1	rpart
11	CART 2	rpart1SE
12	CART 3	rpart2
13	CART or Ordinal Responses	rpartScore
14	Conditional Inference Random Forest	cforest
15	Conditional Inference Tree 1	ctree
16	Conditional Inference Tree 2	ctree2
17	Continuation Ratio Model for Ordinal Data	vglmContRatio
18	Cost-Sensitive C5.0	C5.0Cost
19	Cost-Sensitive CART	rpartCost
20	DeepBoost	deepboost
21	eXtreme Gradient Boosting 1	xgbDART
22	eXtreme Gradient Boosting 2	xgbTree
23	Flexible Discriminant Analysis	fda
24	Generalized Linear Model	glm
25	Generalized Linear Model with Stepwise Feature Selection	glmStepAIC
26	Model Averaged Neural Network	avNNet
27	Multivariate Adaptive Regression Spline	earth
28	Multivariate Adaptive Regression Splines	gcvEarth
29	Neural Network	nnet
30	Neural Networks with Feature Extraction	pcaNNet
31	Penalized Discriminant Analysis	pda
32	Penalized Discriminant Analysis	pda2
33	Penalized Multinomial Regression	multinom
34	Random Forest	ranger
35	Single C5.0 Ruleset	C5.0Rules
36	Single C5.0 Tree	C5.0Tree
37	Stochastic Gradient Boosting	gbm
38	Tree Models from Genetic Algorithms	evtree
39	Bagged AdaBoost	AdaBag
40	Ensembles of Generalized Linear Models	randomGLM
41	Parallel Random Forest	parRF
42	Random Ferns	rFerns

43	Random Forest	rf
44	Random Forest by Randomization	extraTrees
45	Random Forest Rule-Based Model	rfRules
46	Regularized Random Forest	RRF
47	Regularized Random Forest	RRFglobal
48	Weighted Subspace Random Forest	wsrf
49	Bayesian Additive Regression Trees	bartMachine
50	Naive Bayes	naive_bayes
51	Naive Bayes	nb
52	AdaBoost Classification Trees	adaboost
53	AdaBoost.M1	AdaBoost.M1
54	Boosted Classification Trees	ada
55	Boosted Linear Model	BstLm
56	Boosted Logistic Regression	LogitBoost
57	Boosted Tree	bstTree
58	eXtreme Gradient Boosting	xgbLinear
59	L2 Regularized Linear Support Vector Machines with Class Weights	svmLinearWeights2
60	Linear Support Vector Machines with Class Weights	svmLinearWeights
61	Support Vector Machines with Class Weights	svmRadialWeights
62	Distance Weighted Discrimination with Polynomial Kernel	dwdPoly
63	Distance Weighted Discrimination with Radial Basis Function Kernel	dwdRadial
64	Factor-Based Linear Discriminant Analysis	RFlda
65	Heteroscedastic Discriminant Analysis	hda
66	High Dimensional Discriminant Analysis	hdda
67	Linear Discriminant Analysis	lda
68	Linear Discriminant Analysis	lda2
69	Linear Discriminant Analysis with Stepwise Feature Selection	stepLDA
70	Linear Distance Weighted Discrimination	dwdLinear
71	Localized Linear Discriminant Analysis	loclda
72	Maximum Uncertainty Linear Discriminant Analysis	Mlda
73	Mixture Discriminant Analysis	mda
74	Quadratic Discriminant Analysis	qda
75	Quadratic Discriminant Analysis with Stepwise Feature Selection	stepQDA
76	Regularized Discriminant Analysis	rda
77	Robust Regularized Linear Discriminant Analysis	rrlda
78	Shrinkage Discriminant Analysis	sda
79	Sparse Linear Discriminant Analysis	sparseLDA
80	Sparse Mixture Discriminant Analysis	smda
81	Stabilized Linear Discriminant Analysis	slda
82	Sparse Distance Weighted Discrimination	sdwd
83	Rotation Forest	rotationForest
84	Rotation Forest	rotationForestCp
85	Tree-Based Ensembles	nodeHarvest
86	Partial Least Squares	kernelpls
87	Partial Least Squares	pls
88	Partial Least Squares	simpls
89	Partial Least Squares	widekernelpls
90	Sparse Partial Least Squares	spls
91	Gaussian Process	gausssprLinear
92	Gaussian Process with Polynomial Kernel	gausssprPoly
93	Gaussian Process with Radial Basis Function Kernel	gausssprRadial
94	Generalized Additive Model using LOESS	gamLoess
95	Generalized Additive Model using Splines	bam

96	Generalized Additive Model using Splines	gam
97	glmnet	glmnet
98	Multi-Step Adaptive MCP-Net	msaenet
99	Penalized Ordinal Regression	ordinalNet
100	C4.5-like Trees	J48
101	Logistic Model Trees	LMT
102	Nearest Shrunken Centroids	pam
103	Rule-Based Classifier	JRip
104	Rule-Based Classifier	PART
105	Single Rule Classification	OneR
106	L2 Regularized Support Vector Machine (dual) with Linear Kernel	svmLinear3
107	Least Squares Support Vector Machine with Radial Basis Function Kernel	lssvmRadial
108	Support Vector Machines with Linear Kernel	svmLinear
109	Support Vector Machines with Linear Kernel	svmLinear2
110	Support Vector Machines with Polynomial Kernel	svmPoly
111	Support Vector Machines with Radial Basis Function Kernel	svmRadial
112	Support Vector Machines with Radial Basis Function Kernel	svmRadialCost
113	Support Vector Machines with Radial Basis Function Kernel	svmRadialSigma
114	Regularized Logistic Regression	regLogistic
115	Multi-Layer Perceptron	mlpWeightDecay
116	Multi-Layer Perceptron, multiple layers	mlpWeightDecayML
117	Penalized Logistic Regression	plr
118	Radial Basis Function Network	rbfDDA
119	Robust SIMCA	RSimca
120	Monotone Multi-Layer Perceptron Neural Network	monmlp
121	Multi-Layer Perceptron	mlp
122	Multi-Layer Perceptron, with multiple layers	mlpML
123	Stacked AutoEncoder Deep Neural Network	dnn
124	Partial Least Squares Generalized Linear Models	plsRglm
125	Patient Rule Induction Method	PRIM
126	Greedy Prototype Selection	protoclass
127	k-Nearest Neighbors	kknn
128	k-Nearest Neighbors	knn
129	Learning Vector Quantization	lvq
130	Optimal Weighted Nearest Neighbor Classifier	ownn
131	Stabilized Nearest Neighbor Classifier	snn
132	SIMCA	CSimca
133	ROC-Based Classifier	rocc
134	Fuzzy Rules Using Chi's Method	FRBCS.CHI
135	Fuzzy Rules with Weight Factor	FRBCS.W
136	Self-Organizing Maps	xyf