On the evaluation of example-dependent cost-sensitive models for tax debts classification

Helton Souza Lima[1], **Damires Yluska de Souza Fernandes**[1], **Thiago José Moura**[1]

[1]Federal Institute of Paraíba, 720 Avenida Primeiro de Maior, João Pessoa, Brazil

helton.souza@academico.ifpb.edu.br, {damires, thiago.moura}@ifpb.edu.br

Abstract. *Example-dependent cost-sensitive classification methods are suitable to many real-world classification problems, where the costs, due to misclassification, vary among every example of a dataset. Tax administration applications are included in this segment of problems, since they deal with different values involved in the tax payments. To help matters, this work presents an experimental evaluation which aims to verify whether cost-sensitive learning algorithms are more cost-effective on average than traditional ones. This task is accomplished in a tax administration application domain, what implies the need of a cost-matrix regarding debt values. The obtained results show that cost-sensitive methods avoid situations like erroneously granting a request with a debt involving millions of reals. Considering the savings score, the cost-sensitive classification methods achieved higher results than their traditional method versions.*

1. Introduction

Tax revenue is one of the most necessary financial resources of a government for accomplishing specific goals such as health services, mobility, security, education, among others [19]. Failure to comply with tax obligations may have a negative impact on the quality of life of citizens. Tax administrations, whether federal or state, have the control of tax evasion as well as the mitigation of nonpayment risks among their main activities. To help matters, these sectors have tried to gather historical data in order to provide services enhanced by predictive analytics models. The objective underlying these models is usually to identify tax evasion or even taxpayers with a high risk profile for failure to meet financial obligations [23].

Particularly in Brazil, the country's National Treasury Attorney-General's Office (hereafter called as *PGFN*, which has been abbreviated from "Procuradoria Geral da Fazenda Nacional") has aimed to achieve predictive models in order to support decisions on their services. One of the most important services at PGFN is named as "Request for Revision of Registered Debt" (hereafter called as R3D). R3D is an online service, that allows taxpayers to request a reanalysis of the situation of their identified debts.

An R3D can be registered by his/her taxpayer, for instance, in cases of debt payment, administrative decision, correction of statement, or any kinds of extinction or suspension cause. Each request is then analysed by the PGFN in order to decide on its acceptance or rejection. Nowadays, the referred analysis is completely human-dependent and time-consuming, since it makes use of many personal and infrastructure resources. Once the request is granted, the debt's registration may be canceled or rectified. If the

request is rejected, the debt remains valid. In this case, the taxpayer can either pay the debt or proceed to a tax enforcement process.

Not only to support human decisions on the processes but also to reduce the costs of debt analyses, thus enabling faster and more assertive decisions, some supervised classification models have been evaluated to indicate the likelihood for an R3D to be approved or rejected [16]. [16] evaluated some cost-insensitive learning algorithms regarding important and traditional measures w.r.t. the problem of provinding predictions on R3Ds.

Cost-insensitive learning algorithms (or traditional classifiers) assume that all misclassification errors, resulting from type I (false positives) and type II (false negatives) errors, carry the same cost [12]. Still, this is not the case in many real-world applications. For example, failing in an approval of a loan to a fraudster leads to higher losses than denying it to a borrower in good faith. Learning methods that take different misclassification costs into account are known as cost-sensitive classifiers [10, 26]. Regarding the R3D classification problem, cost-insensitive classification models may take the risk of granting a request involving millions of Brazilian reals in debts that should not be forgiven.

With this scenario in mind, we define the main research question that has guided this work: Are cost-sensitive learning algorithms more cost-effective on average than traditional ones when dealing with the R3D classification problem?

We have accomplished an experimental evaluation in order to answer our research question. To this end, this work compares some cost-sensitive classification methods applied to the R3D classification problem w.r.t. their respective versions of cost-insensitive methods. The evaluation is assessed by means of not only traditional metrics such as *accuracy*, *recall*, *precision* and *f-score*, but also by a specific cost *savings* score. The *savings* score is a cost-sensitive evaluation measure, which is defined in accordance with business-oriented rules. Thus, it has been established as the most important measure to be considered in this comparative work.

The main contributions of this work are as follows: (i) definition of a cost-matrix to the R3D classification problem; (ii) an experimental evaluation accomplished to compare cost-sensitive classification methods with their traditional classification method versions; (iii) a comparison among different example-dependent cost-sensitive methods in the light of the R3D classification problem.

This paper is organized as follows: Section 2 provides some theoretical background; Section 3 discusses some related works; Section 4 describes the experimental protocol, the used dataset and the defined cost-matrix; Section 5 presents the results which have been obtained, and Section 6 concludes the paper and suggests some future work.

2. Theoretical Background

Some concepts w.r.t. cost-sensitive classification methods are introduced as follows.

2.1. Cost-insensitive and cost-sensitive classification

Classification algorithms predict qualitative values, which will be assigned in predefined categories [22]. For example, in this work, we deal with a binary classification problem. In the light of our business-oriented problem, an example (instance) is positive in case of a rejected R3D (request) while negative examples regard accepted requests.

Traditionally, cost-insensitive learning algorithms, focus on maximizing *accuracy*, and assume that costs for misclassification errors remain equal [21]. The most used metrics to evaluate the performance of classification methods consider only the number of these misclassification errors [12]. These metrics are usually the following ones [12]: *accuracy* (ACC), *precision* (PRE), *recall* (REC) and *f-score* (F1).

Recently, a significant level of attention has also been paid to cost-sensitive learning algorithms [10]. This is due to the fact that, in many real-world applications, the costs of false positives and false negatives are different. For example, in medical classification problems, predicting that a sick patient is healthy is generally a more serious error than predicting that a healthy patient is sick. Classification methods that use different misclassification costs are known as cost-sensitive classifiers [10].

A classification problem is said to be class-dependent cost-sensitive if costs are different among classes but constant among examples [10, 15]. Typical cost-sensitive approaches assume a constant cost for each type of error, in the sense that the cost depends on the class and it is the same among examples.

A class-dependent approach may not be suitable for many real-world applications. For example, in credit card fraud detection, a fraudulent transaction may implicate a small or big economic value. With respect to the R3D classification problem, debts range from thousands of reals to billions of reals. An error in accepting a request with a smaller debt is less costly than an error in accepting a request with a debt involving millions of reals. An example-dependent cost-sensitive classification problem occurs when costs are not constant for all the examples.

An example-dependent approach can be differentiated from a class-dependent approach in the definition of the cost matrix, as illustrated in Table 1. For the class-dependent approach the costs of each misclassification error are constant for every example: C_{FP} = cost of a false positive; C_{FN} = cost of a false negative; C_{TN} = cost of a true negative and C_{TP} = cost of a true positive. For the example-dependent approach, each example *i* carries a different cost: C_{FPi} = cost of a false positive for an instance *i*; C_{FNi} = cost of a false negative for an instance *i*; C_{TNi} = cost of a true negative for an instance *i* and C_{TPi} = cost of a true positive for an instance *i*.

Table 1. Class-dependent vs example-dependent cost matrix.

	Class-dependent		**Example-dependent**	
Predicted/Actual	Positive	Negative	Positive	Negative
Positive	C_{TP}	C_{FP}	C_{TPi}	C_{FPi}
Negative	C_{FN}	C_{TN}	C_{FNi}	C_{TNi}

To build an example-dependent cost-sensitive classification method, the example-dependent cost matrix must be associated with the dataset as a 4-d array attached to each sample of the entire dataset. The following condition has to be fulfilled for every example: $C_{FPi} > C_{TPi}$ and $C_{FNi} > C_{TNi}$, because the costs associated to the incorrect classifications must be greater than the costs associated to the correct classifications.

Example-dependent cost-sensitive classification methods can be grouped according to the phase where the costs are introduced into the solution [4], e.g., (i) during or (ii)

after training. In the former, the changes make the algorithm consider the costs of each example during the training phase to produce cost-sensitive classifiers. In the latter, the cost-sensitive method is applied after the training step of a cost-insensitive classifier.

2.2. Costs during the training phase

Standard impurity metrics, such as misclassification error, entropy or Gini, consider the distribution of classes of each leaf to evaluate the predictive power of a splitting rule, leading to an impurity metric that is based on minimizing the misclassification rate.

An example-dependent cost-sensitive decision tree classifier (ECSDT) considers the costs of each example during the creation of new nodes and pruning of a tree [3]. Instead of using traditional splitting criteria (e.g., Gini), the costs of each tree node are calculated and the gain of using each split is evaluated as the decrease in the total costs of the algorithm. After a tree is built, it is pruned by using a cost-based pruning criterion [3].

The example-dependent cost-sensitive logistic regression (ECSLR) method introduces example-dependent costs by changing the objective function of the model to one that is cost-sensitive [2]. The modification of the objective function uses gradient descent in order to discover the best parameters for the logistic sigmoid function (of the original algorithm) that minimizes the total cost of the model [2].

Ensemble of cost-sensitive decision trees uses example-dependent cost-sensitive decision trees as base learners on random subsamples of the training set. Then it associates them using three different combination methods [4]: the bagging technique (ECSBag) [6] or the Random Forests technique (ECSRF) [8].

Boosting differs from other ensemble learning techniques for training the base learners one after another. Thus, each new estimator tries to correct the errors of its predecessors. AdaBoost [11] is an ensemble learning technique which builds a strong classifier from a weighted vote of multiple weak base learners, usually implemented as a decision stump.

Example-dependent cost-sensitive AdaBoost (ECSAB) considers the cost of each example in the loss function that defines the error of the former classifiers [27]. This loss function defines the sample weight update in each iteration: the examples with a higher cost have a higher weight in the next training iteration. The loss function also defines the importance of each base classifier in the combination step: the classifier that contributes the most to minimize the cost has a higher amount of say in the vote.

2.3. Costs after training an algorithm

The Bayes Minimum Risk (BMR) method is a post-processing method that converts a cost-insensitive classification algorithm to an example-dependent cost-sensitive classification one. This method consists in quantifying trade-offs among various decisions using probabilities and the costs that accompany such decisions [1]. After a cost-insensitive classifier's training, it takes the estimated probability of each prediction and calculates the risk of predicting each one of the classes considering the misclassification costs. Then it chooses the one with the minimum risk estimated. Considering our R3D classification problem, the BMR method can be defined as in 1 and 2.

$$R(p_a|x) = C(p_a|y_a)P(p_a|x) + C(p_a|y_r)P(p_r|x) \tag{1}$$

$$R(p_a|x) = C(p_a|y_a)P(p_a|x) + C(p_a|y_r)P(p_r|x) \quad (2)$$

where p_a, p_r are the classifier's predictions of accepting or rejecting a R3D, respectively; y_a, y_r are the true labels of accepting or rejecting a R3D; $C(a|b)$ is the cost function when a request is predicted as "a" and the real label is "b"; $P(p_a|x)$, $P(p_r|x)$ are the estimated probabilities for the classifier's prediction for accepting and rejecting a R3D, respectively. Each R3D will be predicted as accepted if $R(p_a|x) \leq R(p_r|x)$.

2.4. An example-dependent cost-sensitive evaluation metric

Standard performance metrics, such as *accuracy*, *precision*, *f-score* or *recall*, assume the same cost for the different misclassification errors [12]. Regarding an example-dependent cost-sensitive classification problem, costs of predictions from two classifiers with equal misclassification rates but different numbers of false positive and false negative are not the same, since $C_{FNi} \neq C_{FPi}$.

The *savings* metric considers the costs of each example to compare the performance of different classifiers [4]. Let Z be a set of N examples and each example is associated with their respective costs, represented as $Z_i = [X_i,\ C_{TPi},\ C_{FPi},\ C_{FNi},\ C_{TNi}]$ and a classifier f which predicts label $f(Z_i)$ for each element *i*, then the absolute value of total cost of using f on Z is defined as $C(y, f(Z)) = \sum_{i=1}^{N} C(y_i, f(Z_i))$ [4].

The *savings* score is defined as the total cost of using a classifier versus the cost of using no classifier at all, named base cost. Base cost is the lowest cost of classifying all examples as positive $(f(Z) = 1)$ or negative $(f(Z) = 0)$ and is defined as $C_{base} = min(C(y, 1), C(y, 0))$. The *savings* can be interpreted as the cost improvement of using the classifier under evaluation, and is expressed as in 3. The best classifier is the one with *savings* closer to one.

$$Savings(y, f(Z)) = \frac{C_{base} - C(y, f(Z))}{C_{base}} \quad (3)$$

Considering the R3D classification scenario, the C_{base} is the cost of a classifier that predicts a rejection for all R3D. If the sum of all misclassification costs of a classifier is 0, than the *savings* score will be 1. If the sum of all misclassification costs are higher than 0, than the *savings* score is the equivalent percentage considering the C_{base} cost.

3. Related Works

The works regarding the proposal of new example-dependent cost-sensitive learning methods present comparison evaluations between their respective cost-insensitive versions of the methods [1, 2, 3, 4, 13, 27]. These works use datasets from banks and credit card companies, for applications of fraud detection, credit scoring and direct marketing analysis. In all evaluation comparisons, it is shown that there have been an improvement to the *savings* score when using cost-sensitive methods. On the other hand, it has also been observed a decrease in *accuracy* and *f-score*.

Some works regarding tax administration real-world problems arising from countries, such as Brazil [24, 14], Spain [17] or Italy [5], present the usage of supervised learning methods to assist decision making. These works confirm improvements in decision support applied to tax administration processes. Lima et al. [16] applied cost-insensitive

methods to the R3D classification problem. This work achieved promising results with the cost-insensitive Random Forest model w.r.t. some traditional measures. Concerning works on the tax administration domain, Mehta et al. [20] present an improvement in the *savings* score when identifying tax evasion. The work was done by considering a dataset from the tax department of the Telangana government, India. Example-dependent costs were considered in a modified deep neural network loss function.

Comparing these works with ours, some different aspects are identified as follows. Regarding the works that propose novel cost-sensitive methods, none of them applied the methods on a dataset originated from tax administration systems [1, 2, 3, 4, 13, 27]. The experimental evaluation undertaken in this present work includes a set of methods with the best performances presented in each one of the referred related works. Regarding the works which make use of tax administration datasets, none of them use cost-sensitive classification methods [24, 14, 17, 5]. When taking the R3D classification problem into account [16], the achieved learning model would take the risk of granting a request involving millions of Brazilian reals in debts that should not be forgiven. At last, although Mehta et al., [20] face a tax administration problem using cost-sensitive classification methods, they focus on tax evasion. Tax evasion is a different classification problem compared to the R3D classification problem for bringing up an unbalanced dataset. A difference in the cost matrix is also observed, more specifically, with regards to the true positive case, what causes a possible different behaviour of the models built. In addition, their comparison included only their proposed cost-sensitive method compared to its traditional classifier version and did not include other cost-sensitive methods.

4. Experimental design

The dataset, cost-matrix and experimental setup underlying this work are described.

4.1. Dataset

The dataset has been created from several PGFN data sources, including transactional and analytical systems. The included historical data of the R3Ds encompass the period of November,2018 to March,2022. The dataset has 29 independent variables and 173.709 R3Ds instances. The dataset is not unbalanced: it has a 60/40% proportion between the two classes respectively. Each class represents the analysis result of the request, indicating its approval or rejection.

Personal or business identification information or any feature considered as sensitive were disregarded. For the sake of confidentiality, details regarding the features are not authorized by the PGFN to be detailed in this work. An overview on the dataset features are provided as follows: (i) the request itself (e.g., the request's motivation); (ii) some information describing the debt (e.g., value, age, type, and situation); and (iii) some history of actions and situations associated with the PGFN processes. The value of a debt is the information regarding the costs involved in each possible R3D prediction. We provide a summary of the debts data distribution presented in the PGFN's dataset in Table 2.

4.2. A cost matrix to the R3D classification problem

The R3D classification problem consists in predicting if a request should be accepted or rejected. The corresponding cost matrix has been defined in accordance with the PGFN's business rules and with the assistance of some domain experts. It is depicted in Table 3.

Table 2. Summary of debts data distribution.

Count	173.709
Mean	R$ 735.460,00
Standard deviation	R$ 16.325.820,00
1st quartile	less than R$ 3.488,18
2nd quartile	less than R$ 12.778,58
3rd quartile	less than R$ 64.158,57
Maximum	R$ 2.690.038.000,00
Sum of all debts	R$ 127.756.051.982,00

Table 3. Cost matrix for the R3D classification problem.

	Actual rejected	Actual accepted
Predicted rejected	$C_{TPi} = 0$	$C_{FPi} = R\$50.000,00$
Predicted accepted	$C_{FNi} =$ debt value	$C_{TNi} = 0$

Regarding an R3D, the positive class is the rejected result. The true positive case has no associated costs. However, in case of a false positive (a request is predicted to be rejected but actually it should be accepted), the debt would be erroneously ratified. That situation leads to the continuation of the debt collection by the PGFN. This occurs regardless from the debt's value. At PGFN, costs related with misclassifications may include the ones regarding human and infrastructure resources necessary to manage each debt. Cunha et al. [9] gathered data during a research to track and estimate the mean cost of tax enforcement processes. These estimated values have been adjusted to fit our classification problem. The costs have been updated according to inflation matters and, based on PGFN business rules, they have been set to a constant value of R$ 50.000,00.

The negative class is the accepted result. The true negative case has no associated costs. However, in case of a false negative (a request is predicted to be accepted but actually it should be rejected), the debt is forgiven and terminated. The costs of this misclassification is completely dependent on the debt value, as follows: if it contains a high value (greater than R$ 1.000.000,00), it has a high cost; if it contains a low value (lower than R$ 50.000,00), it has a low cost.

The cost matrix complies with the rule: $C_{FPi} > C_{TPi}$ and $C_{FNi} > C_{TNi}$. The false negative is example-dependent. The false positive has a fixed cost, and it is not example-dependent. In some examples, $C_{FPi} \geq C_{FNi}$, but in other ones $C_{FNi} > C_{FPi}$.

4.3. Experimental setup

The main objective of this work is to investigate some cost-sensitive classification methods applied to the R3D classification problem in comparison with their respective traditional classifier versions. We aim to answer the research question defined in Section 1. The comparison takes into account the *savings* score, which is a cost-sensitive evaluation metric. Other traditional metrics such as *accuracy*, *recall*, *precision* and *f-score* are included in order to present the impact on type I error and type II error as well.

The traditional classifiers included in the experimental evaluation comprise: decision tree (DT); logistic regression (LR); ensemble using bagging technique (Bag); ensem-

ble using random forest technique (RF); and ensemble using boosting technique (AB). Each one is verified along with its example-dependent cost-sensitive version, namely: decision tree (ECSDT); logistic regression (ECSLR); ensemble using bagging technique (ECSBag); ensemble using random forest technique (ECSRF); and ensemble using boosting technique, namely Adaboost (ECSAB). At last, each considered cost-insensitive classifier is used as a base learner in conjunction with the BMR method. Experiments have been carried out by means of 10 repetitions of stratified 10-fold cross-validations.

5. Results and discussion

A summary of the obtained results w.r.t. the defined measures is presented in Table 4. Each datapoint in Table 4 therefore corresponds to the average computation value and standard deviation achieved by the classifiers on the provided test data. The best results for each metric is presented in bold. From the analysis of the experimental results, some observations are worth mentioning:

- The first one regards the negative values of the *savings* score obtained by all the traditional (cost-insensitive) classifiers. A negative value of *savings* means that the evaluated classifier has a worse performance considering the achieved costs than a naive classifier: a one that rejects all requests.

- The *savings* score of traditional classifiers also presented a high standard deviation, demonstrating instability when considering the cost-sensitive metric. This behavior occurs because, in some cases, the false negatives involve a high debt value. As huge debt values are rare and might not always be selected to the test dataset, the *savings* score reached high levels in these situations. However, cases of false negatives involving huge debt values cause a high negative impact on the *savings* score. The false negative cases involving high debt values are not acceptable situations in the business domain problem at hand.

- All cost-sensitive classifiers' versions present a higher *savings* score when compared to their traditional learning algorithm version. Thus, we might say that cost-sensitive classification methods indeed have a superior performance measure in terms of cost savings. This means that they are able to contribute significantly with issues from a business real problem when dealing with decisions that involve different misclassification costs.

- ECSDT, ECSLR and ECSBag have achieved very similar scores w.r.t. the five analysed metrics. As the ECSBag is an ensemble composed by several ECSDT, presumably the low variance presented in the ECSDT caused similar votes in the ensemble combination step. The similar performances observed between ECSDT and ECSLR was also reported in other work [3].

- Regarding the ECSAB method, although it did not achieve a high *savings* score, it improved its cost-insensitive learning algorithm version (Adaboost). The ECSAB is an ensemble that uses the cost-insensitive version of the Decision Tree classifier as base estimator, whereas ECSRF and ECSBag are ensembles that use the ECSDT as base estimator. This difference may explain the lowest *savings* score achieved by ECSAB and a higher standard deviation when compared to the other cost-sensitive methods.

The behaviour of the ECSRF method needs a further in-depth study, and probably

Table 4. Mean and Standard Deviation of 10 x 10-fold cross-validation.

Classifier	*Savings*	*F1*	*ACC*	*REC*	*PRE*
DT	-113,05% (±140, 29%)	78,41% (±0, 36%)	82,74% (±0, 28%)	78,48% (±0, 54%)	78,35% (±0, 42%)
DT-BMR	72,29% (±0, 59%)	40,46% (±0, 66%)	62,16% (±0, 38%)	32,18% (±0, 61%)	54,46% (±0, 80%)
ECSDT	61,81% (±0, 52%)	19,77% (±0, 50%)	53,09% (±0, 29%)	14,47% (±0, 41%)	31,19% (±0, 70%)
LR	-548,52% (±174, 96%)	41,09% (±16, 91%)	65,84% (±2, 72%)	32,88% (±14, 12%)	59,89% (±12, 84%)
LR-BMR	61,60% (±0, 68%)	23,43% (±0, 91%)	53,07% (±0, 67%)	17,97% (±0, 70%)	33,646% (±1, 41%)
ECSLR	60,56% (±0, 48%)	19,09% (±0, 61%)	52,27% (±0, 37%)	14,11% (±0, 57%)	29,56% (±0, 66%)
Bag	-121,23% (±136, 31%)	**81,41%** (±0, 35%)	85,75% (±0, 25%)	78,16% (±0, 56%)	84,95% (±0, 43%)
Bag-BMR	80,53% (±0, 60%)	58,46% (±0, 80%)	72,44% (±0, 44%)	48,56% (±0, 88%)	73,44% (±0, 76%)
ECSBag	61,82% (±0, 52%)	19,47% (±0, 53%)	53,23% (±0, 34%)	14,16% (±0, 45%)	31,18% (±0, 74%)
AB	-277,89% (±175, 87%)	66,67% (±0, 41%)	75,04% (±0, 30%)	62,52% (±0, 52%)	71,42% (±0, 49%)
AB-BMR	70,38% (±0, 70%)	36,39% (±0, 90%)	61,24% (±0, 44%)	27,76% (±0, 81%)	52,81% (±1, 04%)
ECSAB	-26,70% (±33, 60%)	65,79% (±0, 19%)	61,15% (±0, 27%)	**94,26%** (±0, 24%)	50,74% (±0, 18%)
RF	-132,09% (±147, 85%)	81,36% (±0, 40%)	**85,82%** (±0, 28%)	77,51% (±0, 58%)	**85,62%** (±0, 43%)
RF-BMR	**80,75%** (±0, 59%)	58,80% (±0, 80%)	72,69% (±0, 44%)	48,80% (±0, 88%)	73,96% (±0, 75%)
ECSRF	19,17% (±24, 05%)	47,77% (±15, 23%)	44,54% (±4, 68%)	74,45% (±32, 99%)	37,98% (±4, 39%)

indicates that the random selection of features produced very different cost-sensitive decision trees, because the standard deviation of *savings* and *f-score* are significantly higher than the other tree-based cost-sensitive methods.

Regarding the usage of the BMR method, it achieved the best *savings* score for each one of the verified classifiers. This also happened even when comparing their results with each one which considered costs during the training phase. By maintaining a low standard deviation, they also seemed to be more stable. The Random Forest classifier used with the BMR method presented the highest savings score of all verified learning methods, achieving 80,75% on *savings*. Reasons underlying these insights include the fact that the BMR method uses the prediction probability of each example to, in some cases, change the prediction from acceptance to rejection or vice versa. Figure 1a shows a graphical

representation of the Random Forest classifier before and after the BMR application. The prediction probabilities are presented on x-axis and the debt values are depicted on y-axis. The picture on the left shows that, if the probability of acceptance is lower than 50%, the request is predicted as rejected (red dots). On the other hand, the request is predicted as accepted if the probability of acceptance is higher than 50% (green dots).

Figure 1b depicts the resulting predictions when the BMR method is applied over Random Forest. As the debt value increases, the risk calculation tends to reject some requests with high debt values involved, even if they have a probability of acceptance higher than 50%. On the other hand, the risk calculation tends to accept some requests with low debt values, even if they have a probability of acceptance lower than 50%. An observed trade-off in cost-sensitive methods regards the decrease of *f-score*, *accuracy*, *recall* and *precision*. This effect is also noticed in other related works [4, 27].

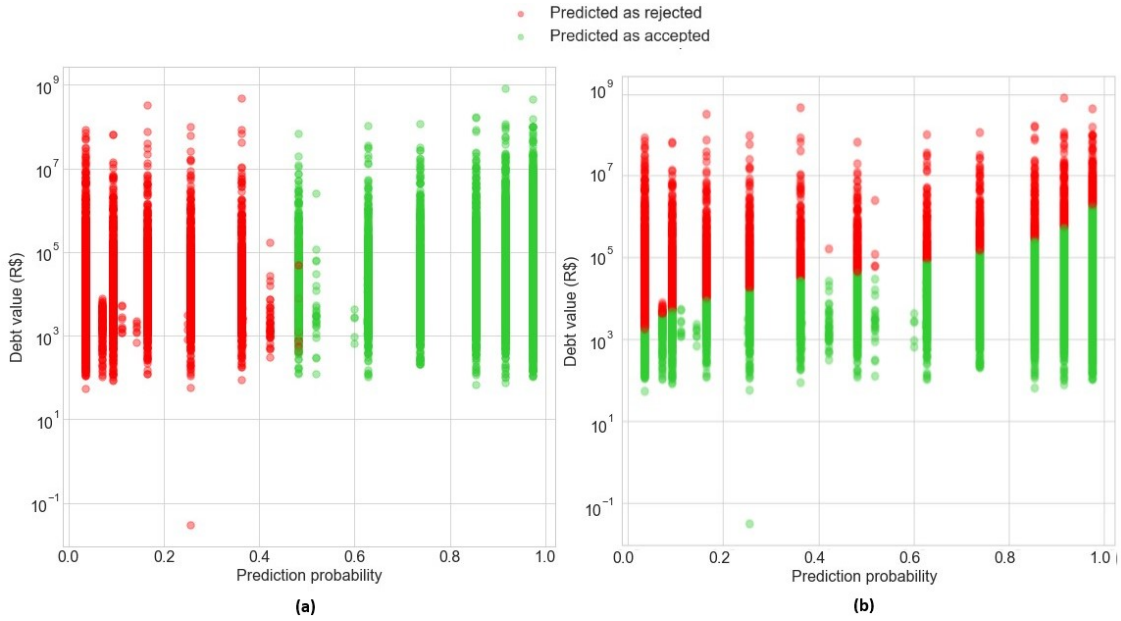

Figure 1. Random Forest's predictions probabilities before (a) and after (b) the application of BMR method.

6. Conclusions and Future Work

This work has presented an experimental evaluation on verifying whether cost-sensitive learning algorithms are more cost-effective on average than traditional ones when dealing with a tax administration classification problem (R3D). To this end, we have defined a cost-matrix in accordance with the business debts value data distribution. This definition is a prerequisite for the usage of any example-dependent cost-sensitive approach.

The experimental results show that traditional classifiers are not the best options when dealing with problems that have different misclassification costs, because they achieve very low levels of a cost-sensitive metric: the *savings* score. It has also been observed high numbers of standard deviation in this same cost-sensitive metric, which means that traditional classifiers are not stable when considering the costs. On the other hand, the cost-sensitive classifiers verified in this work have significantly improved the *savings* score. It means that these classifiers may avoid high financial losses in most of

misclassification cases. Considering the false negative case, that represents a debt that is erroneously forgiven, the severity of this misclassification case depends on the debt amount. If the debt value is low, it has a low cost to the business but; if the debt has a huge value, it has a huge cost to the business and must be avoided.

A remarkable aspect regards the usage of the BMR method applied over the Random Forest classifier (RF-BMR) to the R3D classification problem. In this scenario, we have achieved the best results considering the *savings* score: 80,75%. This result is specially promising to support the analysis of the R3D services, by providing the likelihood for an R3D to be approved or rejected focusing on avoiding expensive costs. As future work we point out some tasks to be done: (i) a further study on the details of the ECSAB and ECSRF behaviours; and (ii) proposition of a new cost-sensitive method applied to the R3D problem that does not present significant increase in false negatives.

References

[1] Bahnsen, A. C., Stojanovic, A., Aouada, D., Ottersten, B.: Cost sensitive credit card fraud detection using Bayes minimum risk. 12th International conference on machine learning and applications (Vol. 1, pp. 333-338). IEEE. (2013)

[2] Bahnsen, A. C., Aouada, D., Ottersten, B.: Example-dependent cost-sensitive logistic regression for credit scoring. 13th International conference on machine learning and applications (pp. 263-269). IEEE. (2014)

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

[4] Bahnsen, A. C., Aouada, D., Ottersten, B.: Ensemble of Example-Dependent Cost-Sensitive Decision Trees. arXiv e-prints, arXiv-1505. (2015)

[5] Battiston, P., Gamba, S., Santoro, A.: Optimizing Tax Administration Policies with Machine Learning. University of Milan Bicocca Department of Economics, Management and Statistics Working Paper, (436). (2020)

[6] Breiman, L.: Bagging predictors. Machine learning, 24(2), 123-140. (1996)

[7] Breiman, L.: Pasting small votes for classification in large databases and on-line. Machine learning, 36(1), 85-103. (1999)

[8] Breiman, L.: Random forests. Machine learning, 45(1), 5-32. (2001)

[9] Cunha, A. D. S., Klin, I. D. V., Pessoa, O. A. G.: Custo e tempo do processo de execução fiscal promovido pela Procuradoria-Geral da Fazenda Nacional. Brasília: Ipea. (2011)

[10] Elkan, C.: The foundations of cost-sensitive learning. International joint conference on artificial intelligence. Vol. 17. No. 1. Lawrence Erlbaum Associates Ltd. (2001)

[11] Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences. (1997)

[12] Harrington, P.: Machine learning in action. 1st edn. Manning Publications. (2012)

[13] Höppner, S., Baesens, B., Verbeke, W., and Verdonck, T.: Instance-dependent cost-sensitive learning for detecting transfer fraud. European Journal of Operational Research, 297(1), 291-300. (2022)

[14] Ippolito, A., Lozano, A. C. G.: Tax Crime Prediction with Machine Learning: A Case Study in the Municipality of São Paulo. In 22nd International Conference on Enterprise Information Systems (pp. 452-459). (2020)

[15] Kim, J., Choi, K., Kim, G., Suh, Y.: Classification cost: An empirical comparison among traditional classifier, Cost-Sensitive Classifier, and MetaCost. Expert Systems with Applications. (2012)

[16] Lima, H. S., de Souza Fernandes, D. Y., Moura, T. J. M., and Sabóia, D.: On the Evaluation of Classification Methods Applied to Requests for Revision of Registered Debts. International Conference on Enterprise Information Systems (ICEIS). (2021)

[17] López, C. P., Rodríguez, M. J. R., Santos, S. L.: Tax fraud detection through neural networks: an application using a sample of personal income taxpayers. Future Internet, 11(4), 86. (2019)

[18] Louppe, G., Geurts, P.: Ensembles on random patches. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 346-361). Springer, Berlin, Heidelberg. (2012)

[19] Mathews, J., Mehta, P., Kuchibhotla, S., Bisht, D., Chintapalli, S. B., Rao, S. K. V.: Regression analysis towards estimating tax evasion in Goods and Services Tax. In IEEE/WIC/ACM International Conference on Web Intelligence. (2018)

[20] Mehta, P., Babu, C. S., Rao, S. K. V., Kumar, S.: DeepCatch: Predicting return defaulters in taxation system using example-dependent cost-sensitive deep neural networks. IEEE International Conference on Big Data (Big Data) (pp. 4412-4419). IEEE. (2020)

[21] Mitchell, T. M.: Machine Learning. McGraw-Hill, 1st edition. (1997)

[22] Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of machine learning. MIT press. (2018)

[23] Ordóñez, P. J., Hallo, M.: Data Mining Techniques Applied in Tax Administrations: A Literature Review. In Sixth International Conference on eDemocracy and eGovernment (ICEDEG) (pp. 224-229). (2019)

[24] Soares, G. V.; Cunha, R. C. L. V.: Predição de Irregularidade Fiscal dos Contribuintes do Tributo ISS. In: Anais do Simpósio Brasileiro de Banco de Dados. (2020)

[25] Wu, R. S., Ou, C. S., Lin, H. Y., Chang, S. I., Yen, D. C.: Using data mining technique to enhance tax evasion detection performance. Expert Systems with Applications, 39(10), 8769-8777. (2012)

[26] Zadrozny, B. , Elkan, C.: Learning and making decisions when costs and probabilities are both unknown. In Proceedings of the seventh ACM SIGKDD inter- national conference on Knowledge discovery and data mining (pp. 204–213). (2001)

[27] Zelenkov, Y. (2019). Example-dependent cost-sensitive adaptive boosting. Expert Systems with Applications, 135, 71-82.

[28] Zhou, Z. H.: Ensemble methods: foundations and algorithms. CRC press. (2012)