# Fraud detection in a two-dimensional decision region under a cost sensitive approach

Jorge C-Rella[1,2*], Ricardo Cao[1†] and Juan M. Vilar[1†]

[1]Department of Mathematics, Research Group MODES, University of A Coruña, CITIC, A Coruña, Spain.

[2]Department of Risks, ABANCA Financial Services, A Coruña, Spain.

*Corresponding author(s). E-mail(s): jorge.crella@udc.es;
Contributing authors: ricardo.cao@udc.es; juan.vilar@udc.es;
†These authors contributed equally to this work.

**Abstract**

Fraud detection is a significantly difficult problem due to the lack of likely patterns, small proportion of positive cases, falsification of data and continuously changing strategies, which causes a recurrent loss in financial institutions. A new approach is proposed on the basis of a loss function, which motivates the construction of an expanded two-dimensional decision space. The expansion allows intrinsically more freedom to the decision region, adjusting it and allowing to adapt it to any restriction. Due to this adaptability an improvement is observed with respect to classical classification techniques. This is proved in a real data set provided by a financial company.

**Keywords:** Fraud detection, Cost-sensitive classification, Decision region, Loss function, Classification

## 1 Introduction

One of the most dangerous risks in a financial company when granting a loan is the credit fraud. Its impact is greater than default risk because, by definition, a fraud is an operation without any payment intention. This leads to the total loss of the financed credit. Furthermore, there are several challenges regarding its detection, which entail a lack of correct surveillance. The first challenge comes due to the lack of likely patterns needed to train any kind of supervised model, caused partly by the second problem, the scarcity of fraudulent cases available. To this is added the constant change in fraud typology. Fraudsters learn from and adapt to risk policies, which makes extremely difficult its categorization. Lastly, there is a lack of reliable data, as fraudsters usually modify or falsify their information which derives in an intrinsic class overlap. All this configures a very adverse setting in order to train a model. If that was not enough, it must be considered that banks can not use complex models as neural networks due to regulatory and implementation restrictions, although they are used in another financial contexts as credit card fraud detection [1, 4, 7, 8].

A few different approaches have been proposed in the financial fraud detection problem. One extended philosophy is to hypothesize that good behavior does not change, so legitimate data points have consistent position in the space and pose the issue as an outlier detection problem [11] or a legitimate subset estimation problem [10]. Low positive class proportion makes fraud probability estimation difficult [6]. Several authors have chosen to tune the estimated probabilities [5, 6], use weighted models [2, 9], apply undersampling techniques [4] or adapt the decision threshold [2, 14, 15]. The main drawback with these approaches is that they do not consider directly the loan amount in the decision making so real losses are not considered.

In practice, the objective is not necessarily to obtain the highest classification accuracy but loss reduction, which is function of the request amount. Thus, this work considers a *loss function* instead of classical metrics that only take into account the classification error probability. With this in mind, instead of focusing on more complex models to estimate the fraud probability, the focus is switched considering a more general decision space taking into account susceptible losses. State of the art models (which will be referred as *classical models* throughout the paper) consider an estimated probability or score, with a subsequent one-dimensional decision region. The proposed method creates an expanded space with the variables that enters the loss function (estimated fraud probability and loan amount in the real problem considered in this paper). Consequently, there is more flexibility and tune possibilities to the decision region.

Next section introduces the problem and the information available for modeling, in order to develop in parallel the explanation of the proposed method and the practical implementation. Section 3 defines the construction of the error measure that motivates the methodology. Section 4 shows the results applying a classical approach, emphasizing on its disadvantages in the cost sensitive setting. Section 5 introduces the construction of the expanded decision space from a logistic regression score and introduces four different proposals. Finally, Section 6 summarizes the computational times and Section 7 the empirical results for all approaches implemented, ending with conclusions and possible extensions.

## 2 Problem description and data available

Most classification algorithms output a probability/score of a data point that measures its likelihood of belonging to the positive class (fraud in this work). In this paper *decision space* is denoted as the probability/score support, where in classical approaches, a cut-off point is selected to create a *decision region* related to the positive class. The problem with this approach in the fraud detection context is twofold. On the one hand it is difficult to calculate an accurate probability estimation due to the small proportion of positive cases and the intrinsic overlap between classes. On the other, although this focus have a good performance when only interest is classification, it have a worse performance in problems where not all kind of errors have the same weight [7, 13]. This is shown in greater depth in Section 4.

In the face of fraud detection, banks have filters and controls for its restraint that lead to a specialist reviewing the operation legitimacy. Here, correct selection of operations to review is important, since not all of them can be reviewed due to the limitation and cost of personnel [1]. In this paper, a restriction related to this is assumed. It consists in having a proportion of transactions to analyze of less than 10%, and ideally smaller than 5%. Taking this into account, operations of greater amount are the most interesting to study, since making a mistake will mean a greater loss.

Problems that address the classification of a binary dependent variable $Y \in \{0, 1\}$ (0 indicating legitimate and 1 fraud) from a set of independent variables $\mathbf{X} = (X_1, \ldots, X_p)$ taking into account costs of prediction error (and potentially other costs) are known as *cost-sensitive problems*. In this setting the two variables of interest are the estimated fraud probability and an exogenous variable, $\xi$, on which the loss depend (the loan amount in the fraud detection problem).

Banks should deny fraudulent transactions at the moment of request, because once the credit is formalized, the money has already been lost. At this moment the only information available is profile data, which could have been manipulated, commerce's historical information within the company and of the request itself. Available variables are aggregated within these three agents. This implies another handicap as the information is limited and a fraudster is not necessarily one with a bad credit profile. Table 1 summarizes the variables used in next sections. It is indicated to which agent each variable refers to, the class and the Information Value (a measure of the relation between a variable and the odds ratio [12]). Due to the limited information available about the client, the commerce takes a key role, given that for him more information is available as it was previously part of the entity's portfolio. Ranges and densities are not shown in order to maintain the portfolio confidentiality of the financial collaborator.

In the practical application a real data set of $210,216$ requests is considered. It was collected between January 2018 to December 2021 with a 0.67% fraud proportion. The number of registers was truncated to change the volume of requests and fraud proportion in order to preserve confidentiality. Only formalized requests are considered, because nothing can be assured about a non-formalized operation. Note that these are the operations of interest (and most difficult to detect) as they are the ones that passed all the filters and controls.

**Table 1** Summary of variables used in the practical implementation. It is indicated which agent each variable corresponds to, its class and its Information Value (IV)

| Variable | Agent | Class | IV |
|---|---|---|---|
| Activity | Commerce | Categorical | 0.227 |
| Activity sector | Client | Categorical | 0.162 |
| Housing situation | Client | Categorical | 0.152 |
| Marital status | Client | Categorical | 0.136 |
| Profession | Client | Categorical | 0.112 |
| Autonomous community | Client | Categorical | 0.108 |
| Class | Commerce | Categorical | 0.032 |
| Previous request indicator | Client | Categorical | 0.007 |
| Monthly amount | Commerce | Continuous | 0.007 |
| Age | Client | Continuous | 0.004 |
| Default rate | Commerce | Continuous | 0.001 |
| Term | Loan | Continuous | 0.001 |

## 3 Loss function

Classical classification techniques evaluate its performance based on a confusion matrix, constructed from the true class $Y$ and the predicted class $\hat{Y} \in \{0,1\}$. Classical evaluation metrics include ROC curves, AUC, mean misclassification error (MME), or accuracy among others. The most widespread approach in the cost-sensitive setting is the consideration of a cost matrix, which assumes that every error of the same type have the same cost. State of the art approaches propose weighted models [9], tuning of the score [5], or selecting the threshold taking into account costs [14], but exogenous variables are never considered explicitly. This means a clear loss of information, for which a loss function is proposed in order to deepen into its effect with a more flexible way to measure the error cost. In addition, an estimation of the expected loss is obtained, the metric that really concerns any business.

In the loss function construction it should be considered all costs/gains associated with all four possibilities arising from the confusion matrix. Extra costs/gains can be considered as well. These include, in order of appearance below, the cost of a perpetrated fraud, the lost gain due to an incorrect dictum of a loan and the personnel cost. Considering $\xi$ the operation amount, the terms in the loss function are given by:

1. $I(Y=1,\hat{Y}=0)\xi$, the total loss of the credit amount due to not detecting the fraud.
2. $c_2c_3I(Y=0,\hat{Y}=1)\xi$, where $c_2$ denotes the analyst error rate (sanctioning a legitimate operation as fraudulent) and $c_3$ the relative percentage mean gain per operation. These values are fixed based on the company experience.
3. $c_1I(\hat{Y}=1)$, where $c_1$ is the mean cost of analyzing a request. Note that this one appears twice in the cost matrix.

As there is no chance of gain when dealing with fraud, only losses are contemplated. Considering all previous cases, the loss function for predicted class $u$, operation amount $t$ and true class $v$ is:

$$\begin{aligned}\ell(u,t,v) = & I(v=1,u=0)t+ \\ & c_2c_3I(v=0,u=1)t+ \\ & c_1I(u=1)\end{aligned} \tag{1}$$

The objective is the minimization of losses and consequently of this function, which is bounded from below by 0 if all frauds are detected with no false positives. In this paper, the accuracy metric considered is *savings*, with an spread use in the literature [1]. It is expressed as:

$$\text{Savings} = 1 - \frac{\sum_{i=1}^{n}\ell(\hat{Y}_i,\xi_i,Y_i)}{\sum_{i=1}^{n}Y_i\xi_i} \tag{2}$$

where the denominator is the total loss faced if no preventive action is taken, and $n$ is the number of operations in the sample.

# 4 One-dimensional decision region

One of the most widespread algorithms used in classification problems is logistic regression. As the objective is to expand and explore the decision space and not necessarily improve estimated probabilities, this model is taken as starting point. Nevertheless it is to be expected that the more accurate the estimated probabilities, the better will be the decision region.

The data set is divided into train and test sets with 70% and 30% of the sample respectively. A logistic model is trained over the train set selecting variables with a stepwise algorithm, considering only significant variables in terms of the $t$-test. Selected ones are summarized in Table 1. Figure 1 summarizes classification metrics calculated over the test set considering different thresholds. In the decision space generated by the estimated probability, $\hat{p} = \hat{P}(Y = 1 \mid \mathbf{X} = \mathbf{x})$, a grid of thresholds is considered dividing this space in 100 segments. Support of the decision space is re-scaled between 0 and 10 in order to preserve confidentiality. From this moment, the scaled estimated probability is denoted as the score $z$.

The foremost highlight about Figure 1 is the uncorrelation between the accuracy and the minimization of the loss function. Accuracy is biased due to the small proportion of frauds, so the maximum is achieved labeling all data points as legitimate. This is one of the main reasons that motivates the consideration of a loss function in this context. The same occur with the percentage of frauds detected. As not all operations have the same amount, there are regions where detecting more frauds worsens the loss function due to a highest increase in false positives. Lastly, note in the top graph the overlap between classes intrinsic to the fraud problem, which makes unlikely to fit a model with high fraud detection and small false positive proportion. Regarding the model itself, it reaches an AUC of 0.751, considered satisfactory taking into account the difficulty of the problem.

In the practical problem, in the test set, although savings can be maximized up to 35.7%, it would imply evaluating 27.1% of the requests, which greatly surpass the imposed restrictions. If positive predicted points thresholds are considered, 29.1% savings can be achieved analyzing 8.4% and 16.3% analyzing 4.8% of the requests. Decision space expansion is considered in the next section. This is expected to improve savings while reducing the proportion of operations to analyze.

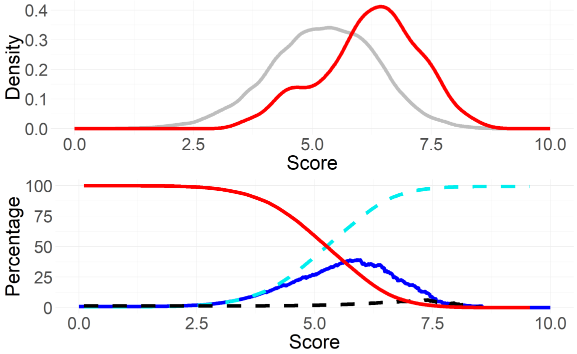

**Fig. 1** Summary graphs for the logistic model considering a grid of thresholds. Top graph represents the score density for legitimate (gray line) and fraudulent (red line) transactions. Bottom graph summarizes various metrics calculated over the test set considering the decision region created by the cut-off point indicated in the horizontal axis. These metrics are the F-score (black dashed line), accuracy (cyan dashed line), percentage of positive predicted points (red solid line) and savings (blue solid line). Classical classification metrics are plotted with a dashed line and the considered in this work with a solid line.

# 5 Two-dimensional decision region

As already noted, the exogenous variable, $\xi$, is not used explicitly by the previous approaches. This section considers expanding the decision space to a two-dimensional map generated by the probability estimation, $\hat{p}$, and the exogenous variable, $\xi$ (loan amount in the fraud detection setting). This approach provides the rudiment to construct a more flexible and effective decision region with a significant impact on the loss function. For instance, a point with a small estimated probability can be classified as fraud if the exogenous variable is high. Note that this is just an extension of the classical decision space, which is of the form $[a, 1] \times \mathbb{R}$ in the expanded decision space. Increasingly adaptable regions are introduced, starting from the least flexible possible (a double cut-off point) to the most one, based on nonparametric techniques.

As will be highlighted later, the smoothness of the decision region have an impact in the performance in the test set. This smoothness is driven by the $k$ parameter, which defines the length of the grid where the optimal search is performed in each proposal. For the two-dimensional sample $\{(z_i, \xi_i)\}_{i=1,\cdots,n}$, a grid denoted by $G(k)$ is defined in the support of the data cloud by a step $\delta_1$ and $\delta_2$ in the first and second dimension respectively:

$$G(K) = \{(z_{\min} + s\delta_1, \xi_{\min} + t\delta_2)\}_{s,t\in\{0,\cdots,k\}} \quad (3)$$

$$\delta_1 = (z_{\max} - z_{\min})/k \quad (4)$$

$$\delta_2 = (\xi_{\max} - \xi_{\min})/k \quad (5)$$

where

$$z_{min} = \min\{z_i\}_{i=1}^n,\ z_{max} = \max\{z_i\}_{i=1}^n,$$
$$\xi_{min} = \min\{\xi_i\}_{i=1}^n,\ \xi_{max} = \max\{\xi_i\}_{i=1}^n$$

## 5.1 Double cut-off point

When considering a two-dimensional decision space, the first idea is to take two cut-off points, one in each dimension, as a generalization of the classical approach introduced in Section 4. Hence the decision region consists of an upper right quadrant in $\mathbb{R}^2$ defined by a cut-off point $\mathbf{x} = (x_1, x_2)$ as $R_{\mathbf{x}} = \{(z, \xi) \in \mathbb{R}^2 \mid z > x_1, \xi > x_2\}$. The optimization is performed evaluating the loss function considering the quadrants generated by the set of points in the grid $G(k)$ with $k = 100$. With this approach, considering the optimal decision region obtained over the train set, savings ascend to 41% for the test dataset, which corresponds to analyzing 26.7% of the requests. This implies an increase of 5.2% in savings with respect to the classical one dimensional decision region estimated in Section 4. The biggest improvement is obtained when considering the 5% positive labeled points restriction, where the difference in savings is of 11% with respect the classical approach. Although the increase in terms of savings is not significant, it implies an improvement with an understandable model. Figure 2 shows the data cloud in the expanded space together with the decision regions obtained under each positive predicted proportion restriction. Note that the most interesting points are those located in the top-right zone, with high probability of being fraud and high amount. This is what triggers the increase in savings compared to the classical approach and is further exploited in next sections.

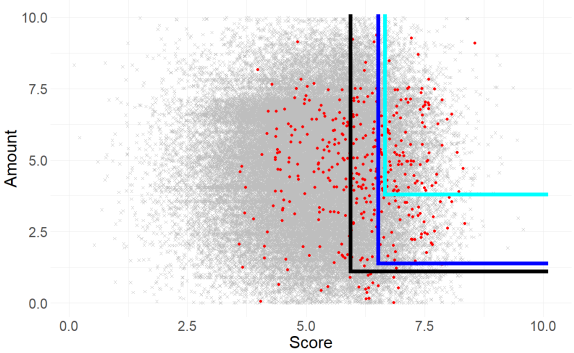

**Fig. 2** Data cloud in the two-dimensional decision space generated by the score and the amount. Legitimate requests are represented with gray crosses and frauds with red dots. Superimposed are represented optimal upper right quadrants in terms of savings, meeting the restrictions of positive labeled points ≤ 100% (black), ≤ 10% (blue), ≤ 5% (cyan).

## 5.2 Bayes minimum risk

The Bayes minimum risk approach [3] is considered as the reference model due to its good practical results and its relation to the methodology proposed. This method considers the exogenous variable in the decision making combined with an estimated probability, $\hat{P}(Y = y \mid \mathbf{X} = \mathbf{x})$ [3]. The model takes into account the *risk* of a data point:

$$\begin{aligned} R(y, \xi \mid x) =& \ell(y, \xi, y)\hat{P}(Y = y \mid \mathbf{X} = \mathbf{x}) + \\ & \ell(y, \xi, 1 - y)(1 - \hat{P}(Y = y \mid \mathbf{X} = \mathbf{x})) \end{aligned}$$

where $y \in \{0, 1\}$ and $\ell$ is the loss function (1). Bayes minimum risk model labels a data point as fraud if $R(1, \xi \mid \mathbf{x}) \leq R(0, \xi \mid \mathbf{x})$, i.e. if the risk of classifying it as a fraud is lower than as legitimate. In the problem under study, considering the loss function defined in (1), for a new data point $i$ this leads to the decision rule:

$$\hat{Y}_i = \begin{cases} 1, & \text{if } \xi_i \geq \dfrac{c_1}{\hat{p}_i(1 + c_2c_3) - c_2c_3} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where $\hat{p}_i = \hat{P}(Y_i = 1 \mid \mathbf{X} = \mathbf{x}_i)$.

Considering the decision region derived from this inequality, in the test set savings of 43.1% are obtained analyzing 26% of the requests. Figure 3 shows the decision region. It is truncated in the graph because as from some point, the minimum amount in the sample is greater than the decision frontier. The main drawback of this approach is that the frontier defined by (6) does not include any tuning parameter allowing flexibility in order to adapt the decision region to the problem at hand. In the practical problem, the proportion of positive predicted points can not be controlled. So despite its good performance, the method is not flexible enough to fulfill the desired restrictions.

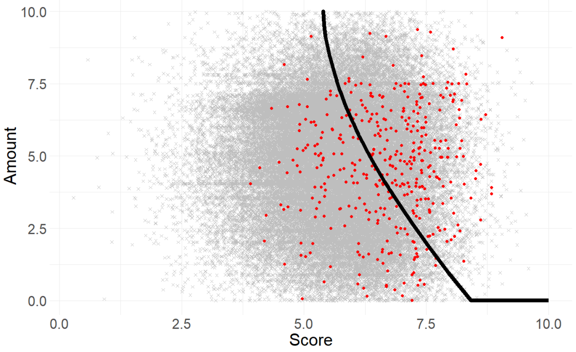

**Fig. 3** Data cloud in the two-dimensional decision space generated by the score and the amount. Legitimate requests are represented with gray crosses and frauds with red dots. Superimposed is represented Bayes minimum risk region.

## 5.3 Quadratic decision region

In this section a parabola is considered for the frontier of the decision region. The loss function is evaluated over the decision region generated by a series of parabolas defined by a set of three parameters and the optimal is selected. The parameters considered for the quadratic function are its vertex (taken from $G(k)$ as defined in (3) with $k = 50$ considering the subset where $z > 5$ and $z < 9$), amplitude (ranging from 0.1 to 3.1 by a step 0.5) and angle of rotation (ranging from 0 to $\Pi/4$ by a step $\Pi/32$). This ranges were set after a preliminary search, in order to reduce the computational time that become boundless if complete ranges were considered. There are configurations of parameters that outputs a parabola that leaves points to its right out of the region, which does not make sense given the problem context. To correct this, in these cases the decision region is extended from the vertex of the parabola to the right of the map, as can be seen in Figure 4. For the optimal region, savings in the test set ascend to 42% analyzing 26% of the transactions, which improves the double threshold approach in terms of both savings and percentage of points inside the region. In order to exploit the susceptible improvements when considering a more flexible region than a single quadrant, next section introduces a nonparametric approach for the decision region estimation.

In addition, optimal parabolas fulfilling any restriction on the proportion of positive predicted transactions can be found, just sticking to the subset that satisfies it. Figure 4 displays optimal parabolas for three thresholds on the proportion of positive labeled transactions. The decision region is similar to the Bayes minimum risk approach but leaving out the bottom of the map, which reduces the proportion of fraud flags and improves savings. The major drawback regarding this proposal is the computational time as there is needed an intensive search over the set of parameters and that the optimal decision region does not necessarily have to have a quadratic form in other problems. Next section introduces the most flexible approach proposed, which in addition reduces the computational times.

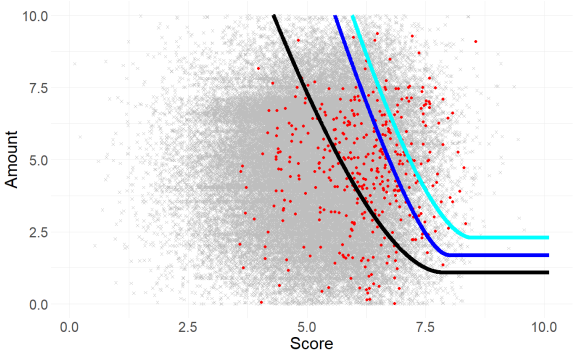

**Fig. 4** Data cloud in the two-dimensional decision space generated by the score and the amount. Legitimate requests are represented with gray crosses and frauds with red dots. Superimposed are represented optimal parabolas in terms of savings, meeting the restrictions of positive labeled points $\leq 100\%$ (black), $\leq 10\%$ (blue), $\leq 5\%$ (cyan).

## 5.4 Nonparametric decision region

The last proposal is a nonparametric approach based on adding quadrants recursively to the decision region until no improvement is found in the loss function. Let's consider $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$, which defines the upper right quadrant $Q_{\mathbf{x}} = \left\{(z, \xi) \in \mathbb{R}^2 \mid z > x_1, \xi > x_2\right\}$. An aggregated decision region defined by a set of points, $R = \{\mathbf{x}_j\}_{j=1,\cdots,r}$, is constructed as the union of their associated upper right quadrants $Q_{\mathbf{x}_j}$,

$$D(R) = \bigcup_{j=1}^{r} Q_{\mathbf{x}_j} \tag{7}$$

Given an observation $(z_i, \xi_i)$, for a decision region $D(R)$ as defined in (7) by a set of frontier points $R$, its predicted indicator of fraud is $\hat{Y}_i^{D(R)} = \mathbb{I}\left((z_i, \xi_i) \in D(R)\right)$. For a sample $\{(z_i, \xi_i)\}_{i=1,\dots,n}$ the value of the loss function considering the decision region $D(R)$ is defined as:

$$\mathcal{L}(R) = \sum_{i=1}^{n} \ell(\hat{Y}_i^{D(R)}, \xi_i, Y_i) \tag{8}$$

Algorithm 1 is proposed for the optimal region estimation. It starts by choosing a single quadrant with vertex in the most northeast point, $(10, 10)$ for our dataset, which corresponds to the one with highest estimated fraud probability and amount. In a recursively manner, each of the points ”surrounding” the current decision region by a step $\delta_1$ as defined in (4) in the first dimension and $\delta_2$ as defined in (5) in the second dimension is added to the current region as in (7) and the associated loss function is calculated as in (8). The point whose inclusion produces the greatest reduction in the loss function is taken. If there is no improvement respect the previous decision region, the process is repeated with a $2\delta_1$, $2\delta_2$ step and so on. Here, the parameter $k$ takes a key role as smoothing parameter. Algorithm stops when the minimum in the data support is reached in the evaluation.

In the non restricted scenario, the proposal consists in just running Algorithm 1. An understandable region is obtained with freedom to select the degree of flexibility taking into account any threshold on the proportion of positive labeled points. For the constrained cases, it is iterated until the restriction (e.g. 10% or 5%) of points inside the region is met. The resulting regions are plotted in Figure 5. Savings in the test set ascend to 43.9% with 24.3% of positive labeled points ($k = 100$), which corresponds to the greatest savings within all the approaches. In the restricted scenarios, depending on the $k$ parameter, the performance in the test set is slightly worse than previous approaches, probably due to a slight over-fitting derived from the restricted search. Regarding the computational time, in Section 6 is commented how empirically it depends quadratically on $k$, which makes this approach suitable for escalation to biggest data sets.

**Algorithm 1** Nonparametric decision region

```
Data  Data set composed by $\mathbf{z}$, $\xi$ and the fraud indicator $\mathbf{Y}$
Input  $k$ parameter
Output  A decision region defined as in (7)
Compute  Grid $G(k)$ as defined in (3);
Preliminary decision region $D(R)$ defined by the point $R = (z_{max}, \xi_{max})$;
Steps $\delta_1$, $\delta_2$ as defined in (4) and (5) respectively;
Frontier $F$ as the subset of $G(k)$ at distance $\delta_1$ and $\delta_2$ in the first and second dimension respectively from $D(R)$
while $\min(F) \geq (z_{max}, \xi_{max})$ do
    $R_{old} \leftarrow R$
    $lf \leftarrow \mathcal{L}(R_{old})$
    $t \leftarrow 1$
    while $R = R_{old}$ do
        Set frontier $F$ as the $G(k)$ subset at distance $t\delta_1$ and $t\delta_2$ in each dimension from $D(R)$
        for Every points in $F$ do
            Compute the loss function as in (8) in the decision region obtained adding the point to $R$ as in (7)
        end for
        Obtain the point $\mathbf{x}_{min} \in F$ whose joining to the current decision region, $R$, outputs the smaller loss function
        Join $\mathbf{x}_{min}$ to $R$
        if $\mathcal{L}(R) \geq lf$ then
            $R \leftarrow R_{old}$
        end if
        $t \leftarrow t + 1$
    end while
end while
```

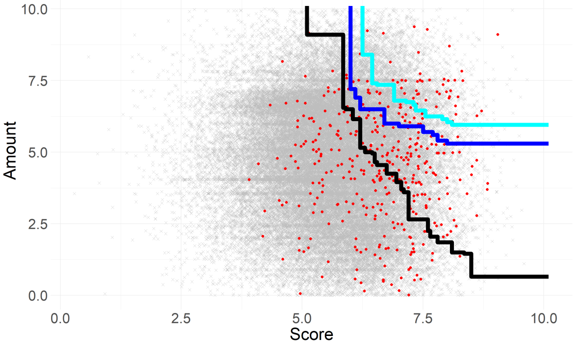

**Fig. 5** Data cloud in the two-dimensional decision space generated by the score and the amount. Legitimate requests are represented with gray crosses and frauds with red dots. Superimposed are represented optimal nonparametric decision regions estimated running Algorithm 1 with $k = 100$ in terms of savings, meeting the restrictions of positive labeled points $\leq 100\%$ (black), $\leq 10\%$ (blue), $\leq 5\%$ (cyan).

# 6 Computational times

Computational times of the different approaches proposed were measured in order to compare them. They are summarized in Table 2 along with the proposals results. Times are considered marginally because in the nonparametric approach, Algorithm 1 is ran and intermediate optimal results satisfying the positive predicted threshold restrictions are saved, so it was only ran one time. The rest of the proposals implies the total evaluation in the grid considered for the posterior optimal selection, which in certain cases lead to a greater computational time than the nonparametric approach. It can be seen empirically how the computational time of Algorithm 1 depends quadratically on the $k$ parameter.

The main advantage of all proposed methods for the approximation of the decision region is that the computational time depends on the size of the search grid and not on the sample size. This allows to scale this methodology to a broad range of problems without necessarily a constraint in the data size. In addition, as computational limitations come from the number of iterations in the optimal search, which can be parallelized or developed in batches, the escalation possibilities are very high. Thus, this methodology fits in any kind of cost-sensitive problem regardless of the sample size and the complexity of the loss function to be optimized.

# 7 Summary and conclusions

This work introduces a new method that can be enriched and refined in so many ways thanks to its flexibility. It is shown how certain regions of the space are not worth considering despite having a high probability of fraud due to the influence of the amount in the loss function. This is exploited with the proposed method. Although three different new approaches are just considered, the are many other possibilities that can be adapted to any problem at hand. It is enough simply to adapt the loss function and the decision space. Regarding the computational costs, as all proposals depend on $k$ and not directly on the sample size as mentioned in Section 6, the method is suitable for its escalation to biggest data sets.

Table 2 summarizes the results of all the approaches introduced throughout the paper. Main highlight is the consistent improvement achieved just expanding the decision space in all proposals compared with the classical approach. Considering Table 2 and the positive predicted proportion threshold mentioned in Section 2, the best choice consists in the nonparametric approach with the 5% restriction. This decision region makes an easy-to-understand rule, satisfying the strictest restriction and with a significant good performance in terms of loss reduction.

**Supplementary information.** ‘Not applicable’

**Acknowledgments.** This research has been supported by MICINN Grant PID2020-113578RB-I00 and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14 and Centro de Investigación del Sistema Universitario de Galicia ED431G 2019/01), all of them through the European Regional Development Fund (ERDF). The first author was financed by the Axencia Galega de Innovación Grant 14-IN606D-2021-2607768.

# Statements and Declarations

**Competing Interests.** ‘Not applicable’

**Consent for publication.** Abanca Servicios Financieros consents to the publication of this work after verifying that no type of confidential or sensitive information related to the company or its clients is provided.

**Table 2** Summary table of the metrics of interest (savings defined in (1) and percentage of positive predicted points) for the proposed approaches throughout the paper for the train sample (top) and the test sample (bottom). Last column displays the computational time in minutes for the construction of each decision region.

| TRAIN set | Unrestricted | | <10% | | <5% | | |
|---|---|---|---|---|---|---|---|
| Decision region | Savings | % | Savings | % | Savings | % | Time (minutes) |
| One dimensional | 41.22 | 27.34 | 25.05 | 8.53 | 18.00 | 4.85 | 0.12 |
| Double threshold | 43.48 | 27.16 | 29.44 | 9.63 | 22.66 | 4.88 | 14.19 |
| Bayesian | 45.79 | 26.23 | | | | | 1.01 |
| Quadratic | 46.64 | 26.21 | 36.79 | 9.80 | 27.54 | 4.92 | 6475.95 |
| Nonparametric ($k = 20$) | 46.91 | 25.12 | 32.53 | 9.03 | 25.64 | 4.36 | 1.24 |
| Nonparametric ($k = 50$) | 47.91 | 24.03 | 36.19 | 9.91 | 29.10 | 4.44 | 26.29 |
| Nonparametric ($k = 100$) | 48.65 | 24.63 | 37.57 | 9.64 | 29.86 | 4.61 | 176.23 |
| TEST set | Unrestricted | | <10% | | <5% | | |
| Decision region | Savings | % | Savings | % | Savings | % | Time (minutes) |
| One dimensional | 35.74 | 27.10 | 29.09 | 8.36 | 16.26 | 4.75 | 0.12 |
| Double threshold | 40.98 | 26.72 | 33.24 | 9.32 | 27.31 | 4.65 | 14.19 |
| Bayesian | 43.06 | 25.99 | | | | | 1.01 |
| Quadratic | 41.96 | 25.95 | 32.72 | 9.50 | 26.58 | 4.75 | 6475.95 |
| Nonparametric ($k = 20$) | 43.41 | 24.85 | 27.31 | 9.01 | 21.87 | 4.22 | 1.24 |
| Nonparametric ($k = 50$) | 43.85 | 23.73 | 30.03 | 9.76 | 27.55 | 4.29 | 26.29 |
| Nonparametric ($k = 100$) | 43.92 | 24.35 | 31.61 | 9.36 | 25.59 | 4.38 | 176.23 |

# References

[1] Almhaithawi, D., Jafar, A., and Aljnidi, M. (2020). Example-dependent cost-sensitive credit cards fraud detection using SMOTE and Bayes minimum risk. *SN Applied Sciences*, 2(1574).

[2] Bahnsen, A. C., Aouada, D., and Ottersten, B. (2014). Example-dependent cost-sensitive logistic regression for credit scoring.

[3] Bahnsen, A. C., Stojanovic, A., Aouada, D., and Ottersten, B. (2013). Cost sensitive credit card fraud detection using Bayes minimum risk.

[4] Dal Pozzolo, A., Caelen, O., Johnson, R. A., and Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. IEEE Symposium Series on Computational Intelligence, 159-166.

[5] Elkan, C. (2001). The foundations of cost-sensitive learning. *Proceedings of the Seventeenth International Conference on Artificial Intelligence*, 1.

[6] King, G. and Zeng, L. (2002). Logistic regression in rare events data. *Political Analysis*, 9.

[7] Lucas, Y. and Jurgovsky, J. (2020). Credit card fraud detection using machine learning: A survey.

[8] Omar, S., Kiwanuka, F., and Swaib, K. (2018). A state-of-the-art review of machine learning techniques for fraud detection.

[9] Pesantez-Narvaez, J. and Guillen, M. (2020). Penalized logistic regression to improve predictive capacity of rare events in surveys. *Journal of Intelligent & Fuzzy Systems*, 38:1–11.

[10] Porwal, U. and Mukund, S. (2018). Credit card fraud detection in e-commerce: An outlier detection approach.

[11] Ramírez, A., Ochoa-Zezzatti, A., Padilla, A., and Ponce, J. (2011). Outlier analysis for plastic card fraud detection a hybridized and multi-objective approach.

[12] Siddiqi, N. (2006). *Credit risk scorecards, developing and implementing intelligent credit scoring*. John Wiley & Sons, Inc., Hoboken, NJ.

[13] Wang, H., Kou, G., and Peng, Y. (2020). Multi-class misclassification cost matrix for credit ratings in peer-to-peer lending. *Journal of the Operational Research Society*, 72:1–12.

[14] Yih, W.-t., Goodman, J., and Hulten, G. (2006). Learning at low false positive rates.

[15] Zadrozny, B. and Elkan, C. (2001). Learning and making decisions when costs and probabilities are both unknown. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.