APPLICATION NOTE

Reject Inference Methods in Credit Scoring

Adrien Ehrhardt[a, b, c] and Christophe Biernacki[b, c] and Vincent Vandewalle[b, d] and Philippe Heinrich[c] and Sébastien Beben[e]

[a]Groupe Crédit Agricole, Groupe de Recherche Opérationnelle, Montrouge, France; [b]Inria; [c]Université de Lille, Laboratoire Paul Painlevé, Villeneuve d'Ascq, France; [d]ULR 2694 - METRICS : Évaluation des technologies de santé et des pratiques médicales, F-59000 Lille, France; [e]BNP Paribas Personal Finance, Levallois-Perret, France.

ARTICLE HISTORY
Compiled May 23, 2021

ABSTRACT

The granting process of all credit institutions is based on the probability that the applicant will refund his/her loan given his/her characteristics. This probability also called score is learnt based on a dataset in which rejected applicants are *de facto* excluded. This implies that the population on which the score is used will be different from the learning population. Thus, this biased learning can have consequences on the scorecard's relevance. Many methods dubbed "reject inference" have been developed in order to try to exploit the data available from the rejected applicants to build the score. However most of these methods are considered from an empirical point of view, and there is some lack of formalization of the assumptions that are really made, and of the theoretical properties that can be expected. In order to propose a formalization of such usually hidden assumptions for some of the most common reject inference methods, we rely on the general missing data modelling paradigm. It reveals that hidden modelling is mostly incomplete, thus prohibiting to compare existing methods within the general model selection mechanism (except by financing "non-fundable" applicants, which is rarely performed in practice). So, we are reduced to empirically assess performance of the methods in some controlled situations involving both some simulated data and some real data (from Crédit Agricole Consumer Finance (CACF), a major European loan issuer). Unsurprisingly, no method seems uniformly dominant. Both these theoretical and empirical results not only reinforce the idea to carefully use the classical reject inference methods but also to invest in future research works for designing model-based reject inference methods, which allow rigorous selection methods (without financing "non-fundable" applicants).

KEYWORDS
reject inference, credit risk, scoring, data augmentation, scorecard, semi-supervised learning

1. Introduction

1.1. Aim of reject inference

For a new applicant's profile and credit's characteristics, the lender aims at estimating the repayment probability. To this end, the *credit modeller* fits a predictive model, often a logistic regression, between already financed clients'

Corresponding Author: Adrien Ehrhardt. Email: adrien.ehrhardt@credit-agricole-sa.fr

characteristics $\boldsymbol{x} = (x_1, \ldots, x_d)$, here d characteristics, and their repayment status, a binary variable $y \in \{0, 1\}$ (where 1 corresponds to "good" clients and 0 to "bad" clients). The model is then applied to the new applicant and yields an estimate of its repayment probability, called score after an increasing transformation (*e.g.* the logit in the case of logistic regression). Over some cut-off value of the score, the applicant is accepted, except if further "expert" rules (*e.g.* credit bureau information, overindebtedness) or an operator come into play.

The through-the-door population (all applicants) can be classified into two categories thanks to a binary variable z taking values in $\{\mathrm{f}, \mathrm{nf}\}$ where f stands for financed applicants and nf for non-financed ones. As the repayment variable y is missing for non-financed applicants, credit scorecards are only constructed on financed clients' data but then applied to the whole through-the-door population. The relevance of this process is a natural question which lies at the heart of *reject inference*. The idea is to use the characteristics of non-financed clients in the scorecard building process to avoid a population bias, and thus to improve the prediction on the whole through-the-door population. Such methods have been described in [2, 9, 18, 23] among others.

1.2. Literature review

Formalization of the reject inference problem is of first importance given the potential financial stakes for credit organizations we previously mentioned. It has notably been investigated in [8] who first saw reject inference as a missing data problem. More precisely, it can be addressed as a part of the semi-supervised learning setting, which consists in learning from both labelled and unlabelled data. However, in the semi-supervised setting, it is generally assumed that labelled data and unlabelled data come from the same distribution (see Ref. [4]), which is rarely the case in *Credit Scoring*. Note that the case of a global misspecified model (both for labelled and unlabelled data), addressed by the initial work in [11], can also complicate this concern. Moreover, the main use case of semi-supervised learning is when the number of unlabelled data is far larger than the number of labelled data, which is not the case in *Credit Scoring* since the number of rejected clients and accepted clients is often balanced and depends heavily on the financial institution, the portfolio considered, *etc.* Consequently, reject inference and related methods require specific studies.

Recent papers (see Refs. [12], [10], [1], [13], [15], [20], [25]) proposed reject inference techniques for other models than the usual logistic regression, but still relying on predictive modelling. Many of them can be cast into the general framework which we will introduce in Section 3 (see also specific discussion of Section 3.9). Beyond their (lack of) prior interpretation and/or justification, most of these proposed models are also difficult to validate *a posteriori*. Indeed, conclusions of the related papers rely either on numerical experiments related to financed clients only (with eventually sample weights depending on the outcome of the loan), or on a pre-processing step inferring the status of non-financed clients by the proposed model itself. Consequently, since the access to a true test set gathering both financed and non-financed clients is not allowed, it becomes impossible to fairly compare such methods (see discussion on such a model selection in Section 2.5).

Alternatively, Mancisidor *et al.* [15] proposes a generative modelling strategy. Indeed, estimating the joint distribution $p(\boldsymbol{x}, y)$ can be straightforwardly applied to partially labelled data through the semi-supervised principle. However, it requires stronger hypotheses on the data-generating mechanism which can unfortunately lead to worse results than predictive models if modelling bias is too high, except if this latter is controlled through information criteria (see Section 2.5), which is not really performed in most literature of this domain.

Ultimately, all these reject inference methods bear the same two major flaws: they are heuristics with implicit hypotheses and without theoretical guarantees; they cannot be empirically evaluated either since experiments always rely on biased samples. This paper aims to bridge these two gaps with some theoretical arguments, illustrated also through selected (simulated and real) numerical experiments.

1.3. Outline of the paper

The purpose of the present paper is thus to revisit most widespread reject inference methods in order to clarify which mathematical hypotheses, if any, underlie these heuristics. This rational review is a fundamental step for raising clear conclusions on their relevance. The question of retaining a reject inference method has also to be addressed in a formal way, namely in the general model selection paradigm.

The outline of the paper is the following. In Section 2, we recast the reject inference concern as a missing data problem embedded in a general parametric modelling. It allows to discuss the related missing data mechanisms, a standard likelihood-based estimation process and also some possible model selection strategies. In Section 3, the most common reject inference methods are described and their mathematical properties are exhibited. These latter mostly rely on the missing data framework previously introduced in Section 2. However, we show that such a theoretical understanding cannot assess the expected quality of the score provided by each method. Subsequently, each method is empirically tested and compared on simulated and real data from CACF in Section 4 to illustrate that no method is universally superior. Finally, some guidelines are given to both practitioners (when using existing reject inference methods) and statisticians (when designing new reject inference methods) in Section 5.

2. *Credit Scoring* modelling

2.1. Data

The decision process of financial institutions to accept a credit application is easily embedded in the probabilistic framework. The latter offers rigorous tools for taking into account both the variability of applicants and the uncertainty on their ability to pay back the loan. In this context, the important term is $p(y|\boldsymbol{x})$, designating the probability that a new applicant (described by his characteristics $\boldsymbol{x}$) will pay back his loan ($y = 1$) or not ($y = 0$). Estimating $p(y|\boldsymbol{x})$ is thus an essential task of any *Credit Scoring* process.

To perform estimation, a specific n-sample (the observed sample) $\mathcal{T}$ is available, decomposed into two disjoint and meaningful subsets, denoted by $\mathcal{T}_{\text{f}}$ and $\mathcal{T}_{\text{nf}}$ ($\mathcal{T} = \mathcal{T}_{\text{f}} \cup \mathcal{T}_{\text{nf}}$, $\mathcal{T}_{\text{f}} \cap \mathcal{T}_{\text{nf}} = \emptyset$). The first subset ($\mathcal{T}_{\text{f}}$) corresponds to applicants $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})$, described by d features, who have been financed ($z_i = \text{f}$) and, consequently, for whom the repayment status y_i is known. With their respective notation $\mathbf{x}_{\text{f}} = \{\boldsymbol{x}_i\}_{i\in\text{F}}$, $\mathbf{y}_{\text{f}} = \{y_i\}_{i\in\text{F}}$ and $\mathbf{z}_{\text{f}} = \{z_i\}_{i\in\text{F}}$, where $\text{F} = \{i : z_i = \text{f}\}$ denotes the corresponding subset of indexes, we have thus $\mathcal{T}_{\text{f}} = \{\mathbf{x}_{\text{f}}, \mathbf{y}_{\text{f}}, \mathbf{z}_{\text{f}}\}$. The second subset ($\mathcal{T}_{\text{nf}}$) corresponds to other applicants $\boldsymbol{x}_i$ who have *not* been financed ($z_i = \text{nf}$) and, consequently, for whom the repayment status y_i is *unknown.* With their respective notation $\mathbf{x}_{\text{nf}} = \{\boldsymbol{x}_i\}_{i\in\text{NF}}$, $\mathbf{y}_{\text{nf}} = \{y_i\}_{i\in\text{NF}}$ and $\mathbf{z}_{\text{nf}} = \{z_i\}_{i\in\text{NF}}$, where $\text{NF} = \{i : z_i = \text{nf}\}$ denotes the corresponding subset of indexes, we have thus $\mathcal{T}_{\text{nf}} = \{\mathbf{x}_{\text{nf}}, \mathbf{z}_{\text{nf}}\}$. We notice that y_i values ($i \in \text{NF}$) are excluded from the observed sample $\mathcal{T}_{\text{nf}}$, since they are missing. Finally, the following notation will be also used: $\mathbf{x} = \{\mathbf{x}_{\text{f}}, \mathbf{x}_{\text{nf}}\}$.

It should be noticed that we use the “financed” versus “not financed” terminology whereas most previous work use “accepted” versus “rejected” clients. Indeed, these two concepts are different: one might be accepted, but never return the contract and / or supporting documents, thus being not financed and yielding a missing label y (this client might have had a better offer elsewhere). Also, a “rejected” client, be it by the score, or specific rules, might be (manually) financed by an operator, who might have had “proof” that the client is good. In these two cases, the common assumption that rejected clients would be performing worse than accepted ones, all else being equal, is false. As discussed in Section 3.7, these kinds of unverifiable assumptions fail to generalize from one financial institution to another. We thus make no distinction inside the “not financed” population in what follows.

2.2. General parametric model

Estimation of $p(y|\boldsymbol{x})$ has to rely on modelling since the true probability distribution is unknown. Firstly, it is both convenient and realistic to assume that triplets in the complete sample $\mathcal{T}_{\text{c}} = \{\boldsymbol{x}_i, y_i, z_i\}_{1 \leq i \leq n}$ are all independent and identically distributed (i.i.d.), including the unknown values of y_i when $i \in \text{NF}$. Secondly, it is usual and convenient to assume that the unknown distribution $p(y|\boldsymbol{x})$ belongs to a given parametric family $\{p_{\boldsymbol{\theta}}(y|\boldsymbol{x})\}_{\boldsymbol{\theta} \in \Theta}$, where Θ is the parameter space. For instance, logistic regression is often considered in practice, even if we will be more general in this section. However, logistic regression will be important for other sections since some standard reject inference methods are specific to this family (Section 3) and numerical experiments (Section 4) will implement them.

As in any missing data situation (here z indicates if y is observed or not), the relative modelling process, namely $p(z|\boldsymbol{x}, y)$, has also to be clarified. For convenience, we can also consider a parametric family $\{p_{\boldsymbol{\phi}}(z|\boldsymbol{x}, y)\}_{\boldsymbol{\phi} \in \Phi}$, where $\boldsymbol{\phi}$ denotes the parameter and Φ the associated parameter space of the financing mechanism. Note that we consider here the most general missing data situation, namely a Missing Not At Random (MNAR) mechanism (see Ref. [14]). It means that z can be stochastically dependent on some missing data y, *i.e.* $p(z|\boldsymbol{x}, y) \neq p(z|\boldsymbol{x})$. We will discuss this fact in Section 2.4.

Finally, combining both previous distributions $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ and $p_{\boldsymbol{\phi}}(z|\boldsymbol{x}, y)$ leads to express the joint distribution of (y, z) conditionally to $\boldsymbol{x}$ as:

$$p_{\boldsymbol{\gamma}}(y, z|\boldsymbol{x}) = p_{\boldsymbol{\phi}(\boldsymbol{\gamma})}(z|y, \boldsymbol{x}) p_{\boldsymbol{\theta}(\boldsymbol{\gamma})}(y|\boldsymbol{x}) \tag{1}$$

where $\{p_{\boldsymbol{\gamma}}(y, z|\boldsymbol{x})\}_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}}$ denotes a distribution family indexed by a parameter $\boldsymbol{\gamma}$ evolving in a space $\boldsymbol{\Gamma}$. Here it is clearly expressed that both parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ can depend on $\boldsymbol{\gamma}$, even if in the following we will note shortly $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\gamma})$ and $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\gamma})$. In this very general missing data situation, the missing process is said to be *non-ignorable*, meaning that parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ can be functionally dependent (thus $\boldsymbol{\gamma} \neq (\boldsymbol{\phi}, \boldsymbol{\theta})$). We also discuss this fact in Section 2.4.

2.3. Maximum likelihood estimation

Mixing previous model and data, the maximum likelihood (ML) principle can be invoked for estimating the whole parameter $\boldsymbol{\gamma}$, thus yielding as a by-product an estimate of the parameter $\boldsymbol{\theta}$. Indeed, $\boldsymbol{\theta}$ is of particular interest, the goal of the financial institutions being solely to obtain an estimate of $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$. The observed log-likelihood can be written as:

$$\ell(\boldsymbol{\gamma}; \mathcal{T}) = \sum_{i \in \text{F}} \ln p_{\boldsymbol{\gamma}}(y_i, \text{f}|\boldsymbol{x}_i) + \sum_{i' \in \text{NF}} \ln \left[\sum_{y \in \{0,1\}} p_{\boldsymbol{\gamma}}(y, \text{nf}|\boldsymbol{x}_{i'}) \right]. \tag{2}$$

Within this missing data paradigm, the Expectation-Maximization (EM) algorithm (see Ref. [5]) can be used: it aims at maximizing the expectation of the complete likelihood $\ell_c(\boldsymbol{\gamma}; \mathcal{T}_{\text{c}})$ (defined hereafter) over the missing labels. Starting from an initial value $\boldsymbol{\gamma}^{(0)}$, iteration (s) of the algorithm is decomposed into the following two classical steps:

E-step: compute the conditional probabilities of missing y_i values ($i \in \text{NF}$):

$$y_i^{(s)} = p_{\boldsymbol{\theta}(\boldsymbol{\gamma}^{(s-1)})}(1|\boldsymbol{x}_i, \text{nf}) = \frac{p_{\boldsymbol{\gamma}^{(s-1)}}(1, \text{nf}\,|\boldsymbol{x}_i)}{\sum_{y' \in \{0,1\}} p_{\boldsymbol{\gamma}^{(s-1)}}(y', \text{nf}\,|\boldsymbol{x}_i)}; \tag{3}$$

M-step: maximize the conditional expectation of the complete log-likelihood:

$$\ell_c(\boldsymbol{\gamma}; \mathcal{T}_{\text{c}}) = \sum_{i=1}^{n} \ln p_{\boldsymbol{\gamma}}(y_i, z_i|\boldsymbol{x}_i) = \sum_{i \in \text{F}} \ln p_{\boldsymbol{\gamma}}(y_i, \text{f}\,|\boldsymbol{x}_i) + \sum_{i \in \text{NF}} \ln p_{\boldsymbol{\gamma}}(y_{i'}, \text{nf}\,|\boldsymbol{x}_{i'}), \tag{4}$$

leading to:

$$\begin{aligned}\boldsymbol{\gamma}^{(s)} &= \underset{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}}{\operatorname{argmax}}\, \mathbb{E}_{\mathbf{y}_{\mathrm{nf}}}[\ell_c(\boldsymbol{\gamma};\mathcal{T}_{\mathrm{c}})|\mathcal{T},\boldsymbol{\gamma}^{(s-1)}] \\ &= \underset{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}}{\operatorname{argmax}} \sum_{i\in\mathrm{F}} \ln p_{\boldsymbol{\gamma}}(y_i,\mathrm{f}\,|\boldsymbol{x}_i) + \sum_{i'\in\mathrm{NF}} \sum_{y\in\{0,1\}} y_{i'}^{(s)} \ln p_{\boldsymbol{\gamma}}(y,\mathrm{nf}\,|\boldsymbol{x}_{i'}).\end{aligned}$$

Usually, stopping rules rely either on a predefined number of iterations, or on a predefined stability criterion of the observed log-likelihood.

2.4. Some current restrictive missingness mechanisms

The latter parametric family is very general since it considers both that the missingness mechanism is MNAR and non-ignorable. But in practice, it is common to consider ignorable models for the sake of simplicity, meaning that $\boldsymbol{\gamma} = (\boldsymbol{\phi},\boldsymbol{\theta})$. Missingness mechanisms and ignorability are more formally defined in Section 1 in the Supplemental Material. There exists also some restrictions to the MNAR mechanism.

The first restriction to MNAR is the Missing Completely At Random (MCAR) setting, meaning that $p(z|\boldsymbol{x},y) = p(z)$. In that case, applicants should be accepted or rejected without taking into account their descriptors $\boldsymbol{x}$. Such a process is not realistic at all for representing the actual process followed by financial institutions. Consequently it is always discarded in *Credit Scoring*.

The second restriction to MNAR is the Missing At Random (MAR) setting, meaning that $p(z|\boldsymbol{x},y) = p(z|\boldsymbol{x})$. The MAR missingness mechanism seems realistic for *Credit Scoring* applications, for example when financing is based solely on a function of $\boldsymbol{x}$, *e.g.* in the case of a score associated to a cut-off, provided all clients' characteristics of this existing score are included in $\boldsymbol{x}$. It is a usual assumption in *Credit Scoring* even if, in practice, the financing mechanism may depend also on unobserved features (thus not present in $\boldsymbol{x}$), which is particularly true when an operator adds a subjective, often intangible, expertise. In the MAR situation the log-likelihood (2) can be reduced to:

$$\ell(\boldsymbol{\gamma};\mathcal{T}) = \ell(\boldsymbol{\theta};\mathcal{T}_{\mathrm{f}}) + \sum_{i=1}^{n} \ln p_{\boldsymbol{\phi}}(z_i|\boldsymbol{x}_i), \tag{5}$$

with $\ell(\boldsymbol{\theta};\mathcal{T}_{\mathrm{f}}) = \sum_{i\in\mathrm{F}} \ln p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i)$. Combining it with the ignorable assumption, estimation of $\boldsymbol{\theta}$ relies only on the first part $\ell(\boldsymbol{\theta};\mathcal{T}_{\mathrm{f}})$, since the value $\boldsymbol{\phi}$ has no influence on $\boldsymbol{\theta}$. In that case, invoking an EM algorithm due to missing data y is no longer required as will be made explicit in Section 3.2.

2.5. Model selection

At this step, several kinds of parametric model (1) have been assumed. It concerns obviously the parametric family $\{p_{\boldsymbol{\theta}}(y|\boldsymbol{x})\}_{\boldsymbol{\theta}\in\Theta}$, and also the missingness mechanism MAR or MNAR. However, it has to be noticed that MAR versus MNAR cannot be tested since we do not have access to y for non-financed clients (see Ref. [17]). However, model selection is possible by modelling also the whole financing mechanism, namely the family $\{p_{\boldsymbol{\phi}}(z|\boldsymbol{x},y)\}_{\boldsymbol{\phi}\in\Phi}$.

Scoring for credit application can be recast as a semi-supervised classification problem (see Ref. [4] for a thorough reference). In this case, following works in [22], classical model selection criteria can be divided into two categories: either scoring performance criteria as *e.g.* error rate on a test set $\mathcal{T}^{\mathrm{test}}$, or information criteria like *e.g.* the Bayesian Information Criterion (BIC).

In the category of error rate criteria, the typical error rate is expressed as follows:

$$\mathrm{Error}(\mathcal{T}^{\mathrm{test}}) = \frac{1}{|\mathcal{T}^{\mathrm{test}}|} \sum_{i\in\mathcal{T}^{\mathrm{test}}} \mathbb{I}(\hat{y}_i \neq y_i), \tag{6}$$

where $\mathcal{T}^{\text{test}}$ is an i.i.d. test sample from $p(y|\boldsymbol{x})$ and where $\hat{y}_i$ is the estimated value of the related y_i value involved by the estimated model at hand. The model leading to the lowest error value is then retained. However, in the *Credit Scoring* context this criterion family is not available since no sample $\mathcal{T}^{\text{test}}$ is itself available. This problem can be exhibited through the following straightforward expression

$$p(y|\boldsymbol{x}) = \sum_{z \in \{\text{f},\text{nf}\}} p(y|\boldsymbol{x}, z)p(z|\boldsymbol{x}) \tag{7}$$

where $p(y|\boldsymbol{x}, z)$ is unknown and $p(z|\boldsymbol{x})$ is known (to some extent) since this latter is implicitly defined by the financial institution itself. We notice that obtaining a sample from $p(y|\boldsymbol{x})$ would require that the financial institution draws $\mathbf{z}^{\text{test}}$ i.i.d. from $p(z|\boldsymbol{x})$ before observing the results $\mathbf{y}^{\text{test}}$ i.i.d. from $p(y|\boldsymbol{x}, z)$. But in practice it is obviously not the case, a threshold being applied to the distribution $p(z|\boldsymbol{x})$ for retaining only a set of fundable applicants, the non-fundable applicants being definitively discarded, preventing us from getting a test sample $\mathcal{T}^{\text{test}}$ from $p(y|\boldsymbol{x})$. As a matter of fact, only a sample $\mathcal{T}_{\text{f}}^{\text{test}}$ of $p(y|\boldsymbol{x}, \text{f})$ is available, irrevocably prohibiting the calculus of (6) as a model selection criterion.

In the category of information criteria, the BIC criterion is expressed as the following penalization of the maximum log-likelihood:

$$\text{BIC} = -2\ell(\hat{\boldsymbol{\gamma}}; \mathcal{T}) + \dim(\boldsymbol{\Gamma}) \ln n, \tag{8}$$

where $\hat{\boldsymbol{\gamma}}$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\gamma}$ and $\dim(\boldsymbol{\Gamma})$ is the number of parameters to be estimated in the model at hand. The model leading to the lowest BIC value is then retained. Many other BIC-like criteria exist (see Ref. [22]) but the underlined idea is unchanged. Contrary to the error rate criteria like (6), it is thus possible to compare models without funding "non-fundable applicants" since only the available sample $\mathcal{T}$ is required. However, computing (8) requires to precisely express the model families $\{p_{\boldsymbol{\gamma}}(y, z|\boldsymbol{x})\}_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}}$ which compete.

3. Rational reinterpretation of reject inference methods

3.1. The reject inference challenge

As discussed in the previous section, a rigorous way to use the whole observed sample $\mathcal{T}$ in the estimation process implies some challenging modelling and assumption steps. A method using the whole sample $\mathcal{T}$ is traditionally called a reject inference method since it uses not only financed applicants (sample $\mathcal{T}_{\text{f}}$) but also non-financed, or rejected, applicants (sample $\mathcal{T}_{\text{nf}}$). Since modelling the financing mechanism $p(z|\boldsymbol{x}, y)$ is sometimes a too heavy task, such methods propose alternatively to use the whole sample $\mathcal{T}$ in a more empirical manner. However, this is somehow a risky strategy since we have also seen in the previous section that validating methods with error rate like criteria is not possible through the standard *Credit Scoring* process. As a result, some strategies are proposed to perform a "good" score function estimation without access to their real performance, including ignoring non-financed clients as is usually done.

Nevertheless, most of the proposed reject inference strategies have hidden assumptions on the modelling process. Our present challenge is to reveal as far as possible such hidden assumptions to then discuss how realistic these are, if we fail to compare them by the model selection principle.

3.2. Strategy 1: ignoring non-financed clients

3.2.1. Definition

The simplest reject inference strategy is to ignore non-financed clients for estimating $\boldsymbol{\theta}$. Thus it consists in estimating $\boldsymbol{\theta}$ by maximizing the log-likelihood $\ell(\boldsymbol{\theta}; \mathcal{T}_{\text{f}})$.

3.2.2. Missing data reformulation

In fact, this strategy is equivalent to using the whole sample $\mathcal{T}$ (financed and non-financed applicants) under both the MAR and ignorable assumptions. See the related explanation in Section 2.4 and works in [26]. Consequently, this strategy is truly a particular “reject inference” strategy although it does not seem to be.

3.2.3. Estimate property

Denoting by $\hat{\boldsymbol{\theta}}_{\mathrm{f}}$ and $\hat{\boldsymbol{\theta}}$ the MLE of $\ell(\boldsymbol{\theta}; \mathcal{T}_{\mathrm{f}})$ and $\ell_c(\boldsymbol{\theta}; \mathcal{T}_{\mathrm{c}})$, respectively, provided we know y_i for $i \in \mathrm{NF}$, classical ML properties (see Refs. [24] and [26]) yield under a well-specified model hypothesis (there exists $\boldsymbol{\theta}^\star$ s.t. $p(y|\boldsymbol{x}) = p_{\boldsymbol{\theta}^\star}(y|\boldsymbol{x})$ for all $(\boldsymbol{x}, y)$) and an MAR ignorable missingness mechanism that $\hat{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}}_{\mathrm{f}}$ for large enough samples $\mathcal{T}_{\mathrm{f}}$ and $\mathcal{T}$.

3.3. Strategy 2: Fuzzy Augmentation

3.3.1. Definition

This strategy can be found in Ref. [18] and is developed in depth in Section 3.1 in the Supplemental Material. It corresponds to an algorithm which is starting with $\hat{\boldsymbol{\theta}}^{(0)} = \hat{\boldsymbol{\theta}}_{\mathrm{f}}$ (see previous section). Then, all $\{y_i\}_{i \in \mathrm{NF}}$ are imputed by their expected value given by: $\hat{y}_i^{(1)} = p_{\hat{\boldsymbol{\theta}}^{(0)}}(1|\boldsymbol{x}_i)$ (notice that these imputed values are not in $\{0, 1\}$ but in $]0, 1[$). The complete log-likelihood $\ell_c(\boldsymbol{\theta}; \mathcal{T}_{\mathrm{c}}^{(1)})$ given in (4) in a broader context with $\mathcal{T}_{\mathrm{c}}^{(1)} = \mathcal{T} \cup \hat{\mathbf{y}}_{\mathrm{nf}}^{(1)}$ is maximized, with $\hat{\mathbf{y}}_{\mathrm{nf}}^{(1)} = \{\hat{y}_i^{(1)}\}_{i \in \mathrm{NF}}$, and yields final parameter estimate $\hat{\boldsymbol{\theta}}^{(1)}$.

3.3.2. Missing data reformulation

Following the notations introduced in Section 2.3, and recalling that this method does not take into account the financing mechanism $p(z|\boldsymbol{x}, y)$, this method corresponds to a unique iteration of an EM-algorithm yielding $\hat{\boldsymbol{\theta}}^{(1)} = \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{y}_{\mathrm{nf}}}[\ell_c(\boldsymbol{\theta}; \mathcal{T}_{\mathrm{c}}^{(1)})|\mathcal{T}, \hat{\boldsymbol{\theta}}^{(0)}]$. Since $\boldsymbol{\phi}$ is not involved in this process, we first deduce from Section 2.4 that, again, MAR and ignorable assumptions are present.

3.3.3. Estimate property

Some straightforward algebra (solving the M-step of the related EM algorithm) allow to obtain that $\operatorname{argmax}_{\boldsymbol{\theta}} \ell_c(\boldsymbol{\theta}; \mathcal{T}_{\mathrm{c}}^{(1)}) = \hat{\boldsymbol{\theta}}_{\mathrm{f}}$, regardless of any assumption on the missingness mechanism or the true model hypothesis. In other words we have $\hat{\boldsymbol{\theta}}^{(1)} = \hat{\boldsymbol{\theta}}_{\mathrm{f}}$, so that this method is strictly equivalent to the scorecard learnt on the financed clients (Strategy 1 described in Section 3.2).

3.4. Strategy 3: Reclassification

3.4.1. Definition

This strategy corresponds to an algorithm which is starting with $\hat{\boldsymbol{\theta}}^{(0)} = \hat{\boldsymbol{\theta}}_{\mathrm{f}}$ (see Section 3.2). Then, all $\{y_i\}_{i \in \mathrm{NF}}$ are imputed by the *maximum a posteriori* (MAP) principle given by: $\hat{y}_i^{(1)} = \operatorname{argmax}_{y \in \{0,1\}} p_{\hat{\boldsymbol{\theta}}^{(0)}}(y|\boldsymbol{x}_i)$. The complete log-likelihood $\ell_c(\boldsymbol{\theta}; \mathcal{T}_{\mathrm{c}}^{(1)})$ given in (4) in a broader context with $\mathcal{T}_{\mathrm{c}}^{(1)} = \mathcal{T} \cup \hat{\mathbf{y}}_{\mathrm{nf}}^{(1)}$, with $\hat{\mathbf{y}}_{\mathrm{nf}}^{(1)} = \{\hat{y}_i^{(1)}\}_{i \in \mathrm{NF}}$, is maximized and yields parameter estimate $\hat{\boldsymbol{\theta}}^{(1)}$.

Its first variant stops at this value $\hat{\boldsymbol{\theta}}^{(1)}$. Its second variant iterates until potential convergence of the parameter sequence $(\hat{\boldsymbol{\theta}}^{(s)})$, after alternating s iterations between $\hat{\boldsymbol{\theta}}^{(s)}$ and $\hat{\mathbf{y}}_{\mathrm{nf}}^{(s)}$ in a similar way as described above for the first iteration $s = 1$. In practice, and for logistic regression, this method can be found in [9] under the name "Iterative Reclassification", in Ref. [23] under the name "Reclassification" or under the name "Extrapolation" in [2]. It is developed in depth in Section 3.2 in the Supplemental Material.

3.4.2. Missing data reformulation

This algorithm is equivalent to the so-called Classification-EM algorithm where a Classification (or MAP) step is inserted between the Expectation and Maximization steps of an EM algorithm (described in Section 2.3). Classification-EM aims at maximizing the complete log-likelihood $\ell_c(\boldsymbol{\theta}; \mathcal{T}_c)$ over both $\boldsymbol{\theta}$ and $\mathbf{y}_{\text{nf}}$. Since $\boldsymbol{\phi}$ is not involved in this process, we first deduce from Section 2.4 that, again, MAR and ignorable assumptions are present.

3.4.3. Estimate property

Standard properties of the estimate maximizing the complete likelihood indicate that it is not a consistent estimate of $\boldsymbol{\theta}$ according to [3], contrary to the traditional ML one. The related Classification-EM algorithm is also known for "sharpening" the decision boundary: predicted probabilities are closer to 0 and 1 than their true values as is illustrated from simulated data with a MAR ignorable mechanism on Figure 1. The scorecard $\hat{\boldsymbol{\theta}}_{\text{f}}$ on financed clients (in green) is asymptotically consistent as was emphasized in Section 3.2 while the reclassified scorecard (in red) is biased even asymptotically.

We now describe the experimental setup of Figure 1. This setup is similar to the experiment of Section 4.1.1 with $d = 1$ and a single simulated cutpoint: we draw 1 dataset of 10,000 observations of continuous data, homoscedastic and normally distributed s.t. $Y \sim B(\frac{1}{2})$ and $X|Y = y \sim \mathcal{N}(y, 1)$. We learn a logistic regression of coefficient $\hat{\boldsymbol{\theta}}$ and consider that rejected clients ($\mathbf{x}_{\text{nf}}$) correspond to the observations for which $p_{\hat{\boldsymbol{\theta}}}(1|x) < 0.3$. We then learn a logistic regression on "financed" clients only which yields parameter $\hat{\boldsymbol{\theta}}_{\text{f}}$ and one on "reclassified" clients which yields, say, $\hat{\boldsymbol{\theta}}_{\text{CEM}}$. We display on Figure 1 $p_{\hat{\boldsymbol{\theta}}}(1|x)$, $p_{\hat{\boldsymbol{\theta}}_{\text{f}}}(1|x)$ and $p_{\hat{\boldsymbol{\theta}}_{\text{CEM}}}(1|x)$: the "financed" clients model is very close to the "oracle" (the model on the whole population), while the reclassified model is biased and produces a sharper decision boundary.

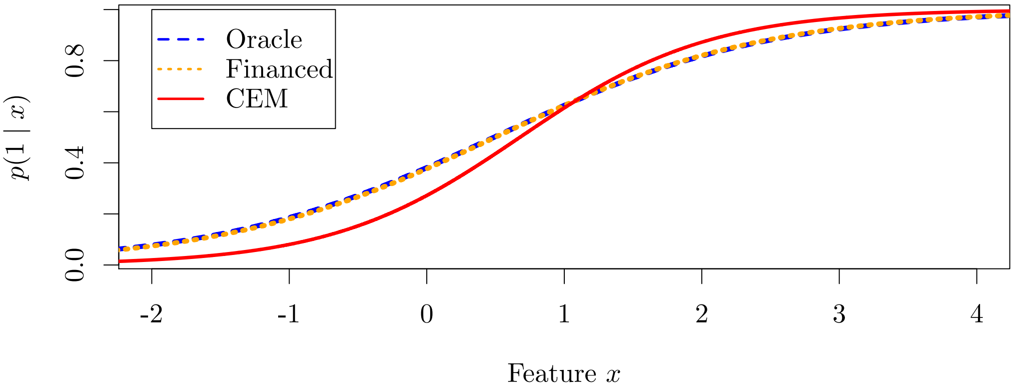

Figure 1. In the context of a probabilistic classifier, it is known that the Classification-EM algorithm employed implicitly by the Reclassification method amounts to a bigger bias in terms of logistic regression parameters, but a "sharper" decision boundary.

3.5. Strategy 4: Augmentation

3.5.1. Definition

Augmentation can be found in [23]. It is also documented as a “Re-Weighting method” in [2, 9, 18] and is described in Section 3.3 in the Supplemental Material. This technique is directly influenced by the importance sampling literature (see Ref. [26] for an introduction in a similar context as here). Indeed, intuitively, as for all selection mechanisms such as survey respondents, observations should be weighted according to their probability of being in the sample w.r.t. the whole population, *i.e.* by $p(z|\boldsymbol{x}, y)$ as was made apparent in Equation (1). By assuming implicitly a MAR and ignorable missingness mechanism, as emphasized in Section 2.4, we get $p(z|\boldsymbol{x}, y) = p(z|\boldsymbol{x})$.

For *Credit Scoring* practitioners, the estimate of interest is the MLE $\hat{\boldsymbol{\theta}}$ of $\ell_c(\boldsymbol{\theta}; \mathcal{T}_\text{c})$, which cannot be computed since we do not know $\mathbf{y}_\text{nf}$. However, recall that the log-likelihood is an empirical criterion derived from maximizing the asymptotic criterion $\mathbb{E}_{\boldsymbol{x},y}[\ln[p_{\boldsymbol{\theta}}(y|\boldsymbol{x})]]$ (the opposite of the Kullback-Leibler divergence between $p(y|\boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ up to a constant w.r.t. $\boldsymbol{\theta}$). By assuming an MAR and ignorable missingness mechanism which leads to $p(y|\boldsymbol{x}, \text{f}) = p(y|\boldsymbol{x})$ for all $\boldsymbol{x}, y$, assuming $p(\text{f}\,|\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$, and noticing that $p(\boldsymbol{x}) = \dfrac{p(\boldsymbol{x}|\,\text{f})}{p(\text{f}\,|\boldsymbol{x})} p(\text{f})$, we can rewrite this asymptotic criterion with the following alternative empirical formulation (valid when $n \to \infty$; recall we denote by $\mathcal{X}$ the space of $\boldsymbol{x}$):

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x},y}[\ln[p_{\boldsymbol{\theta}}(y|\boldsymbol{x})]] &= \sum_{y=0}^{1} \int_{\mathcal{X}} \ln p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) p(y|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \\
&= \sum_{y=0}^{1} \int_{\mathcal{X}} p(\text{f}) \ln p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) \frac{p(\boldsymbol{x}|\,\text{f})}{p(\text{f}\,|\boldsymbol{x})} p(y|\boldsymbol{x}) d\boldsymbol{x} \\
&= p(\text{f}) \sum_{y=0}^{1} \int_{\mathcal{X}} \frac{\ln p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{p(\text{f}\,|\boldsymbol{x})} p(\boldsymbol{x}, y|\,\text{f}) d\boldsymbol{x} \\
&\approx \frac{p(\text{f})}{n} \sum_{i \in \text{F}} \frac{1}{p(\text{f}\,|\boldsymbol{x}_i)} \ln p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i). \qquad (9)
\end{aligned}$$

Recall that the summation over $i \in \text{F}$ means our observations $(\boldsymbol{x}_i, y_i)$ are drawn from $p(\boldsymbol{x}, y|\,\text{f})$, which are precisely the ones we have at hand. Advantage of this new likelihood expression is that, had we access to $p(\text{f}\,|\boldsymbol{x})$, the parameter maximizing this likelihood would asymptotically be equal to the MLE $\hat{\boldsymbol{\theta}}$ maximizing the (complete) log-likelihood $\ell_c(\boldsymbol{\theta}; \mathcal{T}_\text{c})$. However, $p(\text{f}\,|\boldsymbol{x})$ must be estimated by any method retained by the practitioner, which can be a challenging task by itself. In practice, the usual way is to propose to bin observations in $\mathcal{T}$ in K *equal-length* intervals of the score given by $p_{\hat{\boldsymbol{\theta}}_\text{f}}(1|\boldsymbol{x})$ (often $K = 10$) and then to simply estimate $p(z|\boldsymbol{x})$ as the proportion of financed clients in each of these bins. The inverse of this estimate is then used to weight financed clients in $\mathcal{T}_\text{f}$ and finally the score model is retrained within this new context (see again Section 3.3 in the Supplemental Material for all the detailed procedure).

3.5.2. Missing data reformulation

The method aims at correcting for the selection procedure yielding the training data $\mathcal{T}_\text{f}$ in the MAR case. As was argued in Section 3.2, if the model is well-specified, such a procedure is superfluous as the estimated parameter $\hat{\boldsymbol{\theta}}_\text{f}$ is consistent. In the misspecified case however, $\hat{\boldsymbol{\theta}}_\text{f}$ does not converge to the parameter of the best logistic regression approximation $p_{\boldsymbol{\theta}^\star}(y|\boldsymbol{x})$ of $p(y|\boldsymbol{x})$ w.r.t. the aforementioned asymptotic criterion, contrary to the parameter given by this method by construction.

3.5.3. Estimate property

The importance sampling paradigm requires $p(\mathrm{f}\,|\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$, to ensure finite variance of the targeted estimate, which is clearly not the case here: for example, jobless people are never financed. In practice, it is also unclear if the apparent benefit of this method, all assumptions being met, is not offset by the added estimation procedure of $p(\mathrm{f}\,|\boldsymbol{x})$ which remains challenging by itself.

3.6. Strategy 5: Twins

3.6.1. Definition

This reject inference method is documented internally at CACF. It consists in combining two logistic regression-based scorecards: one predicting y learnt on financed clients (denoted by $\hat{\boldsymbol{\theta}}_{\mathrm{f}}$ as previously), the other predicting z learnt on all applicants (denoted by $\hat{\boldsymbol{\phi}}$), before learning the final scorecard using the predictions made by both previous scorecards on financed clients. The detailed procedure is provided Section 3.4 in the Supplemental Material (it is specific to logistic regression).

3.6.2. Missing data reformulation

The method aims at re-injecting information about the financing mechanism in the MAR non-ignorable missingness mechanism by estimating $\hat{\boldsymbol{\phi}}$ as a logistic regression on all applicants, calculating scores $(1,\boldsymbol{x})'\hat{\boldsymbol{\theta}}_{\mathrm{f}}$ and $(1,\boldsymbol{x})'\hat{\boldsymbol{\phi}}$ and use these as two continuous features in a third logistic regression predicting again the repayment feature y, thus using only financed clients in $\mathcal{T}_{\mathrm{f}}$.

3.6.3. Estimate property

From the expression of the log-likelihood (Equation (1) in Section 3.4 in the Supplemental Material), it can be straightforwardly noticed that the logit of $p_{\boldsymbol{\theta}}(y_i|(1,\boldsymbol{x}_i)'\hat{\boldsymbol{\theta}}_{\mathrm{f}},(1,\boldsymbol{x}_i)'\hat{\boldsymbol{\phi}}_{\mathrm{f}})$ is simply a linear combination of $\boldsymbol{x}$, since both $(1,\boldsymbol{x}_i)'\hat{\boldsymbol{\theta}}_{\mathrm{f}}$ and $(1,\boldsymbol{x}_i)'\hat{\boldsymbol{\phi}}_{\mathrm{f}}$ are themselves a linear combination of $\boldsymbol{x}$. Consequently, we strictly obtain $\hat{\boldsymbol{\theta}}^{\mathrm{twins}} = \hat{\boldsymbol{\theta}}_{\mathrm{f}}$. Finally, the last step of the Twins method is known to let the scorecard estimated unchanged (it corresponds to the Fuzzy Augmentation method, see Section 3.3), which allows to conclude that Twins method is strictly identical to the financed clients method given in Section 3.2 (it provides a final scorecard $\hat{\boldsymbol{\theta}}_{\mathrm{f}}$).

3.7. Strategy 6: Parcelling

3.7.1. Definition

The parcelling method can be found in works in [2, 9, 23]. It is also described in Section 3.5 in the Supplemental Material. This method aims to correct the log-likelihood estimation in the MNAR case by making further assumptions on $p(y|\boldsymbol{x},z)$. It is a little deviation from the Fuzzy Augmentation method (Section 3.3) in a MNAR setting, where the payment status $\hat{y}_i^{(1)}$ for non-financed clients ($i \in \mathrm{NF}$) is estimated by a quantity now differing from this one associated to financed clients (which was namely, $p_{\hat{\boldsymbol{\theta}}^{(0)}}(1|\boldsymbol{x}_i,\mathrm{f})$, with $\hat{\boldsymbol{\theta}}^{(0)} = \hat{\boldsymbol{\theta}}_{\mathrm{f}}$). The core idea is to propose an estimate $\hat{y}_i^{(1)} = \hat{p}(1|\boldsymbol{x}_i,\mathrm{nf}) = 1 - \hat{p}(0|\boldsymbol{x}_i,\mathrm{nf})$, for $i \in \mathrm{NF}$, with

$$\hat{p}(0|\boldsymbol{x}_i,\mathrm{nf}) \propto \epsilon_{k(\boldsymbol{x}_i)} p_{\hat{\boldsymbol{\theta}}^{(0)}}(0|\boldsymbol{x}_i,\mathrm{f}),$$

where $k(\boldsymbol{x})$ is the scoreband index among K equal-length scorebands $B_1,\ldots,B_K$ (see Step (b) in Section 3.3 in the Supplemental Material) and $\epsilon_1,\ldots,\epsilon_K$ are so-called “prudence factors”. These latter are generally such that

$1 < \epsilon_1 < \cdots < \epsilon_K$, and they aim to counterbalance the fact that non-financed low refunding probability clients are considered way riskier, all other things being equal, than their financed counterparts. All these ϵ_k values have to be fixed by the practitioner. The method is thereafter strictly equivalent to Fuzzy Reclassification by maximizing over $\boldsymbol{\theta}$ the complete log-likelihood $\ell_c(\boldsymbol{\theta}; \mathcal{T}_{\mathrm{c}}^{(1)})$ with $\mathcal{T}_{\mathrm{c}}^{(1)} = \mathcal{T} \cup \hat{\mathbf{y}}_{\mathrm{nf}}^{(1)}$ and $\hat{\mathbf{y}}_{\mathrm{nf}}^{(1)} = \{\hat{y}_i^{(1)}\}_{i \in \mathrm{NF}}$. It yields a final parameter estimate $\hat{\boldsymbol{\theta}}^{(1)}$.

3.7.2. Missing data reformulation

By considering not-financed clients as riskier than financed clients with the same level of score, *i.e.* $p(0|\boldsymbol{x}, \mathrm{nf}) > p(0|\boldsymbol{x}, \mathrm{f})$, it is implicitly assumed that operators that might have interfered with the system's decision have access to additional information, say $\tilde{\boldsymbol{x}}$ such as supporting documents, that influence the outcome y even when $\boldsymbol{x}$ is accounted for. In this setting, rejected and accepted clients with the same score differ only by $\tilde{\boldsymbol{x}}$, to which we do not have access and is accounted for "quantitatively" in a user-defined prudence factor $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_K)$ stating that rejected clients would have been riskier than accepted ones.

3.7.3. Estimate property

The prudence factor encompasses the practitioner's *belief* about the effectiveness of the operators' rejections. It cannot be estimated from the data nor tested and is consequently a matter of unverifiable expert knowledge.

3.8. Theoretical results summary

Here is provided a compilation of the previous theoretical results:

- Except Strategy 5 (Twins), all strategies are not specific to logistic regression. They can thus be applied to any other parametric scoring method, while not affecting conclusions related to estimate properties.
- Strategies 2 and 5 are equivalent to Strategy 1 consisting in discarding rejected clients $\mathbf{x}_{\mathrm{nf}}$ in the estimation mechanism.
- Strategies 1, 2 and 5 provide consistent estimates of the score function, provided that the missingness mechanism is MAR and ignorable, and providing that the score model is well-specified.
- Strategy 3 (Reclassification) does not provide consistent score estimates, even in the MAR and well-specified cases, since it inherits from the biased properties of the Classification-EM algorithm on which this strategy relies.
- Strategy 4 (Augmentation) is devoted to the misspecified score model case. However, to provide both consistent estimates (related to the oracle candidate score model) and low variance estimates, it requires the user to add likely unavailable additional information through the importance sampling function.
- Strategy 6 (Parcelling) is devoted to the MNAR setting. Again, it requires the user to add likely unavailable additional information about the possible distribution of the MNAR process.

3.9. Other methods related to previous strategies

Previous selected strategies (Strategy 1 to 6) have been detailed since they are representative of many of the used reject inference methods, recent or not. As an example, in recent papers [25], [1], and [13] the "Reclassification" scheme (Strategy 3, Section 3.4) is used in conjunction with LightGBM (and isolation forests for reject inference), Bayesian networks and SVMs respectively, which are global models in the sense of [26]. Such models rely asymptotically on $p(\boldsymbol{x})$ to obtain an estimate of $p(y|\boldsymbol{x})$. In the case of local models such as logistic regression, *i.e.* solely depending on $p(y|\boldsymbol{x})$ asymptotically, we have seen that this strategy produces biased estimates. Hence, aforementioned global models associated with Strategy 3 still produce biased estimates.

More generally, all existing or future reject inference methods *should* be recast from the statistical theory point of view, as we exemplified in this section though their most representative candidates.

4. Numerical experiments

Several authors compared the previously defined reject inference methods through experiments without concluding on what method is best and why, or if so, their results were in contradiction with some other works. For example, it is concluded in [2, 23] that reject inference techniques could not improve credit scorecards. In [9, 18], the opposite is stated. This emphasizes the heuristical nature of these methods, the absence of theoretical guarantees and consequently the fact that the superiority of a method on the others is highly data-dependent. In the previous section, we showed that in theory no reject inference method produces a universally better estimator than the financed clients model.

To support these theoretical findings, we first use simulated data to control under which assumptions (missingness mechanism and well-specified model) we operate. Then we use real data from CACF where we simulate rejected applicants among the financed clients. We exemplify the reject inference methods of the previous section using logistic regression, although our theoretical conclusions apply to all local models. Indeed, logistic regression is still very standard in financial institutions as CACF, provided a satisfactory trade-off between prediction ability and interpretability ability. However, all results are also complemented in Section 4 in the Supplemental Material with other score functions (through the error rate and the Gini value) like decision trees, neural networks and SVMs. Such additional score functions are restricted to Strategy 1 (which is also equivalent to Strategy 2) since devoted to illustrate that we retrieve result variability we already observed in experiments of Section 4.1 and 4.2.

4.1. Simulated data

4.1.1. Results for MAR and well-specified model case

We draw 20 learning and 1 test datasets of 10,000 and 100,000 observations respectively of continuous data, homoscedastic and normally distributed s.t. $Y \sim B(\frac{1}{2})$ and $\boldsymbol{X}|Y = y \sim \mathcal{N}(\mu_y, 2\boldsymbol{I})$ for $y \in \{0, 1\}$ with $\mu_0 = \boldsymbol{0}$, $\mu_1 = \boldsymbol{1}$, and where $\boldsymbol{I}$ denotes the identity matrix. For each learning dataset, we first estimate $\hat{\boldsymbol{\theta}}$ using all observations. Then, we hide y_i by progressively raising the simulated cut-off defining $Z = \text{f}$ if $p_{\hat{\boldsymbol{\theta}}}(1|\boldsymbol{x}_i) > cut$, and $Z = \text{nf}$ otherwise. For each cut-off value *cut* and for each training set, all reject inference methods are trained and we represent their mean Gini (common performance metric in *Credit Scoring* proportional to the Area Under the ROC Curve - higher is better).

This setting is equivalent to a MAR and ignorable missingness mechanism. Logistic regression is well-specified, such that, following our findings in Section 3, we naturally obtained the exact same results from three methods: the logistic regression on financed clients only, the logistic regression using Fuzzy Augmentation and the logistic regression using the Twins method (Strategies 1, 2 and 5 respectively).

We are left with the following four models displayed on Figure 2 and calculated with $d = 8$: the logistic regression on financed clients only, the logistic regression on reclassified data, the logistic regression on augmented data and the logistic regression on parceled data with 10 equal-width score-bands and $\epsilon_k = 1.15$ for $1 \leq k \leq K$ (common in-house practice), corresponding to Strategies 1, 3, 4 and 6 respectively.

What can be concluded from Figure 2 is that logistic regression on financed clients is fine as expected. It is not statistically different from the reclassified dataset and it is significantly better than reject inference using augmented or parceled data as the *cut* becomes larger.

To challenge logistic regression with a "natural" semi-supervised learning approach, we used generative models in the form of Gaussian mixtures (*i.e.* $\boldsymbol{X}|Y = y \sim \mathcal{N}(\mu_y, \boldsymbol{\Sigma}_y)$ - see Ref. [16] for an introduction) for continuous features and multinomial mixtures for categorical ones (see Section 2 in the Supplemental Material for technicalities).

The Gaussian mixture model is not only as good as logistic regression on the left side of Figure 2, all observations being labelled, which was to be expected since it is also well-specified and subsequently benefits from a smaller asymptotical

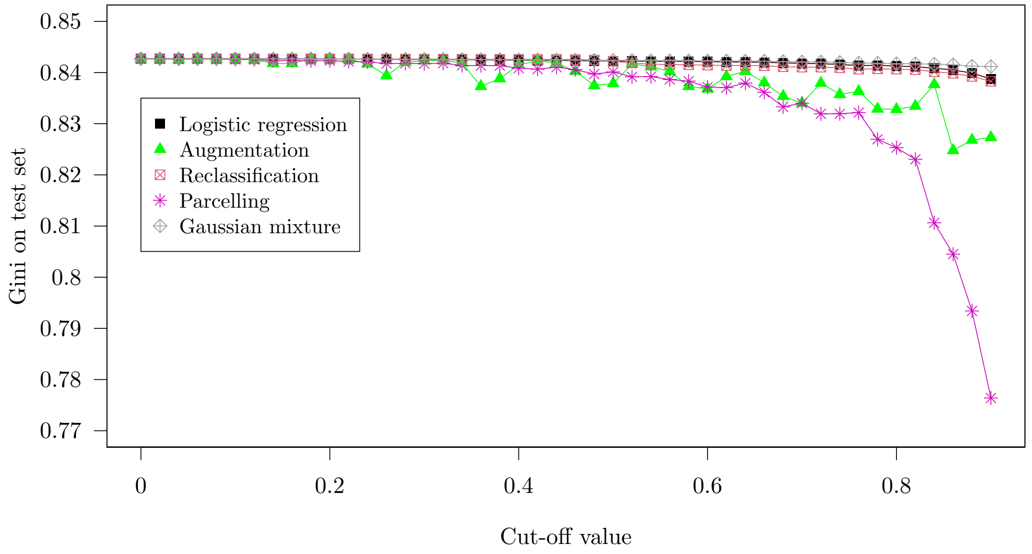

Figure 2. Comparison of reject inference methods with a well-specified model.

variance (see Refs. [6, 19]), but it becomes better than logistic regression-based models when the cut-off becomes larger. This is due to their native use of unlabelled data as detailed in Section 3.

4.1.2. Results for MAR and misspecified model case

In Section 3, we saw that in the MAR and misspecified model case, if some clients beneath the cut-off are accepted so that for all $\boldsymbol{x}$, $p(\mathrm{f}|\boldsymbol{x}) > 0$ then the Augmentation method is well-suited. However, as financing is deterministic here (recall we defined $Z = \mathrm{f}$ if $p_{\hat{\boldsymbol{\theta}}}(1|\boldsymbol{x}_i) > \mathrm{cut}$, and $Z = \mathrm{nf}$ otherwise), this assumption does not hold. To show numerically the consequences of misspecification, we reproduced the same experience as in the previous section, but this time using different variance-covariance matrices for each population (i.e. $\boldsymbol{X}|Y = y \sim \mathcal{N}(\mu_y, \boldsymbol{\Sigma}_y)$ by drawing two random positive definite matrices $\boldsymbol{\Sigma}_y$ for $y \in \{0, 1\}$[1]). In this situation, logistic regression is misspecified whereas the Gaussian mixture remains well-specified. Results are displayed on Figure 3. The gap between logistic regression-based models and the generative model at the beginning of the curve is a clear sign of misspecification for logistic regression. Among logistic regression-based models, the financed clients performed best, the Reclassification method being relatively close. Augmentation and Parcelling fall way behind.

The next section is dedicated to the performance measure of those methods applied on real data from CACF, with which the true model assumption does not hold.

[1] Using the proposed implementation at: `https://stat.ethz.ch/pipermail/r-help/2008-February/153708`

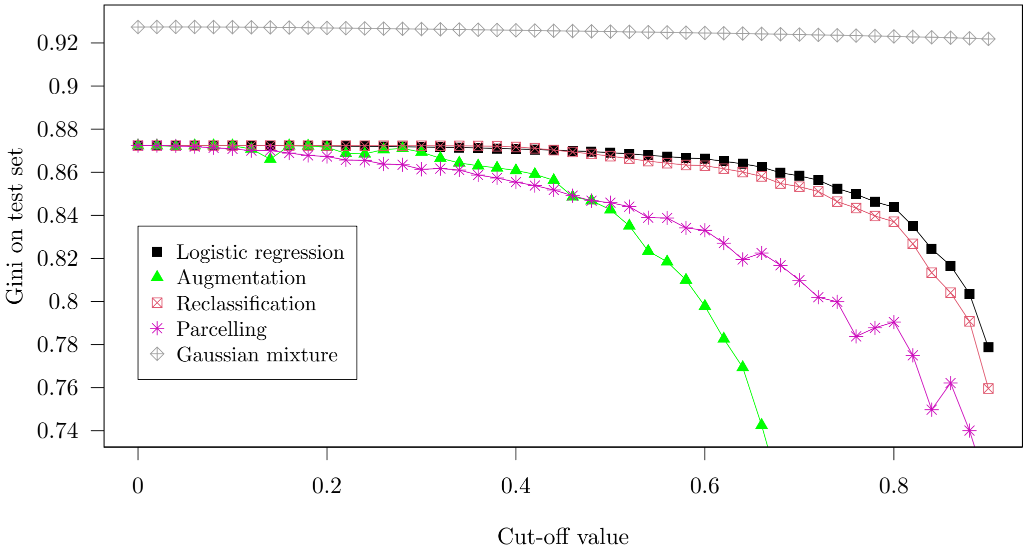

Figure 3. Comparison of reject inference methods with a misspecified model.

4.2. Real data

4.2.1. Procedure

As we have seen up to now and specifically in the preamble of the current section, many authors wrote about reject inference. Those works often had contradictory findings. For authors for which it is necessary, the method to use may differ. This is partly because two learning datasets may give very different results when using a reject inference technique, as we will exhibit here.

We chose three different portfolios in terms of market and product and performed 5-fold cross-validation: we split portfolios so as to learn a model on 80% of the data and compute its Gini on the remaining 20% of the data. Then, we repeat the process and average the Gini measures. As with simulated data, we vary the proportion of simulated financed clients this time by raising the cut-off value of the existing scorecard.

4.2.2. Presentation of the data

The three different portfolios represent financed clients coming from a consumer electronics partner (≈ 200,000 files), a sports company partner (≈ 30,000 files) and the website (≈ 30,000 files) of CACF. Each portfolio has its own existing scorecard. We decided to use the same variables as those already in the scorecards so as to be as close as possible to a MAR process. Those variables are different depending on the scorecard and although we cannot disclose those for confidentiality reasons, they consist in basic information that might come in a scorecard (sometimes crossed with others), *e.g.* the accommodation type (renter, owner, living by family, . . .), the marital status, the age, information on eventual cosigner, *etc.*

Note that there are both categorical and continuous variables. One common practice in the field of *Credit Scoring* is to discretize every continuous variable and group categorical features' levels, if numerous. Again, as we used variables already in the existing scorecard, they are all categorical with 3 to 7 factor levels depending on the variable of interest. Depending on the dataset, there contain approximately 2 to 6% of "bad" clients.

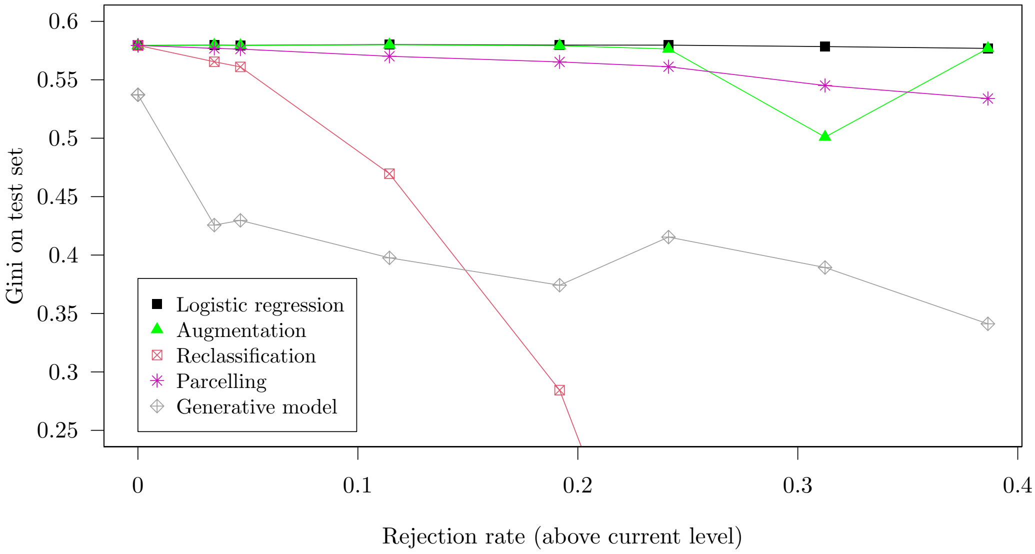

Figure 4. Comparison of several reject inference techniques for the consumer electronics dataset.

4.2.3. Results

Fuzzy Augmentation and Twins had the exact same performance as logistic regression on financed clients, as shown in Section 3, that is why they were excluded from the analysis that follows. Results are displayed on Figures 4, 5 and 6. All logistic regression-based models start at the same Gini because for the first point, there are no rejected applicants. Graphically, it seems that all models produce very similar results for a low cut-off value (left side of the plot) and get bad as soon as it is high. Figures have been voluntarily stopped at approximately 50% acceptance rate due to computational problems: not enough bad clients left in learning set, some categorical features' values not observed anymore in the learning set but present in the test set, *etc.*

The generative approach (a multinomial model here) is not quite as good as logistic regression for low cut-off values most probably for two reasons: first, it makes more assumptions (on $p(\boldsymbol{x})$ - see Ref. [16]) which leads to a greater modelling misspecification; second, the features which were selected were engineered specifically for logistic regression. Nevertheless, its ability to natively use the unlabelled information showed promising results for large cut-off values.

To conclude, Figures 4, 5 and 6 show that no method works significantly and uniformly better than logistic regression on financed clients on all three portfolios.

In previous works, such experiments (often on a single portfolio and a single data point, *i.e.* a single rejection rate) led researchers to conclude positively or negatively on the benefit of reject inference methods. However, given our theoretical findings in Section 3, the fact that in our experiments the results seem highly dependent on the data and/or the proportion of financed clients, the fact that the performance differences were not statistically significant, we shall conclude that reject inference provides no scope for improving the current scorecard construction methodology with logistic regression.

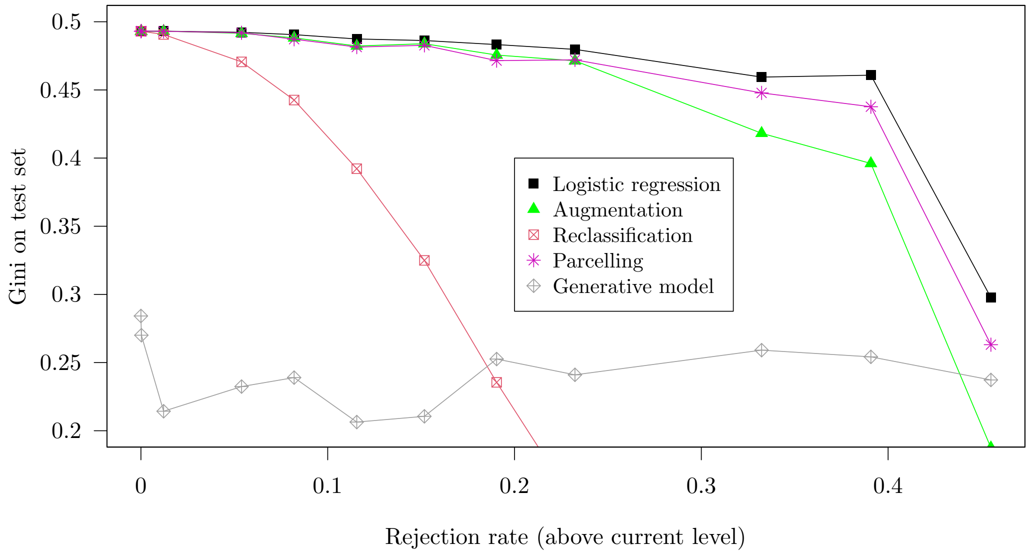

Figure 5. Comparison of several reject inference techniques for the sports company dataset.

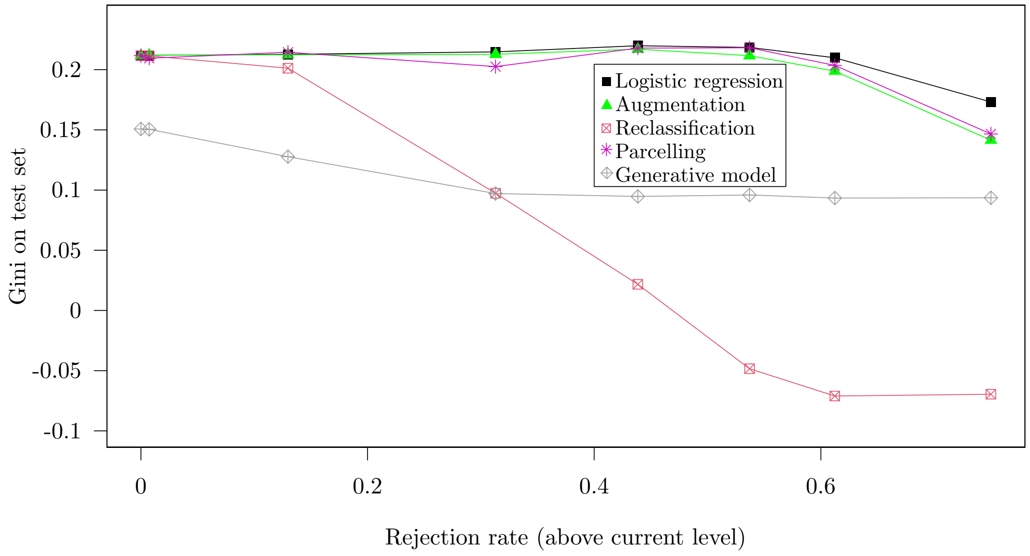

Figure 6. Comparison of several reject inference techniques for the web clients dataset.

5. Discussion: choosing the right model

5.1. Sticking with the financed clients model

Constructing scorecards by using a logistic regression on financed clients is a trade-off: on the one hand, it is implicitly assumed that it is well-specified, and that the missingness mechanism governing the observation of y is MAR and ignorable. In other words, we suppose $p(y|\boldsymbol{x}) = p_{\boldsymbol{\theta}^\star}(y|\boldsymbol{x}, \text{f})$. On the other hand, these assumptions, which seem strong at first hand, cannot really be relaxed: first, the use of logistic regression is a requirement from the financial institution.

Second, the comparison of models cannot be performed using standard techniques since $\mathbf{y}_{\text{nf}}$ is missing (Section 2.5). Third, strategies 4 (Augmentation) and 6 (Parcelling) which tackle the misspecified model and MNAR settings respectively, require additional estimation procedures that, supplemental to their estimation bias and variance, take time from the practitioner's perspective and are rather subjective (see Sections 3.5 and 3.7), which is not ideal in the banking industry since there are auditing processes and model validation teams that might question these practices.

5.2. MCAR through a Control Group

Another simple solution would be to keep a small portion of the population where applicants are not filtered: everyone gets accepted, thus creating a true test set as advocated for in Section 2.5.

Although theoretically perfect, this solution faces a major drawback: it is costly, as many more loans will default. To construct the scorecard, a lot of data is required, so the minimum size of the *Control Group* leads to a much bigger loss than the amount a bank would accept to lose to get a few more Gini points.

5.3. Keep several models in production: "champion challengers"

Several scorecards could also be developed, e.g. one using each reject inference technique. Each application is randomly scored by one of these scorecards. As time goes by, we would be able to put more weight on the most performing scorecard(s) and progressively less on the least performing one(s): this is the field of Reinforcement Learning (see Ref. [21] for a thorough introduction).

The major drawback of this method, although its cost is very limited unlike the *Control Group*, is that it is very time-consuming for the credit modeller who has to develop several scorecards, for the IT who has to put them all into production, for the auditing process and for the regulatory institutions.

6. Concluding remarks

For years, the necessity of reject inference at CACF and other institutions (as it seems from the large literature coverage this research area has had) has been a question of personal belief. Moreover, there even exists contradictory findings in this area.

By formalizing the reject inference problem in Section 2, we were able to pinpoint in which cases the current scorecard construction methodology, using only financed clients' data, could be unsatisfactory: under an MNAR missingness mechanism and/or a misspecified model. We also defined criteria to reinterpret existing reject inference methods and assess their performance in Section 2.5. We concluded that no current reject inference method could enhance the current scorecard construction methodology: only the Augmentation method (Strategy 4) and the Parcelling method (Strategy 6) had theoretical justifications but introduce other estimation procedures. Additionally, they cannot be compared through classical model selection tools (Section 2.5).

We confirmed numerically these findings: given a true model and the MAR assumption, no logistic regression-based reject inference method performed better than the current method. In the misspecified model case, the Augmentation method seemed promising but it introduces a model that also comes with its bias and variance resulting in very close performances compared with the current method. With real data provided by CACF, we showed that all methods gave very similar results: the “best” method (by the Gini) was highly dependent on the data and/or the proportion of unlabelled observations. Last but not least, in practice such a benchmark would not be tractable as $\mathbf{y}_{\text{nf}}$ is missing, thus making it also highly dependent on the way we simulate not-financed clients from previously financed clients. In light of those limitations, adding to the fact that implementing those methods is a non-negligible time-consuming task, we recommend credit modellers to work only with financed loans' data unless there is significant information available on either rejected applicants ($\mathbf{y}_{\text{nf}}$ - credit bureau information for example, which does not apply to France) or on the acceptance mechanism $\boldsymbol{\phi}$ in the MNAR setting. On a side note, it must be emphasized that this work only applies to local models. For global models, explicitly or implicitly obtaining their predictive model $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ as a by-product of modelling $p(\boldsymbol{x})$ or $p(\boldsymbol{x}|y)$, *e.g.* decision trees, it can be shown that they are biased even in the MAR and well-specified settings, thus requiring *ad hoc* reject inference techniques such as an adaptation of the Augmentation method (Strategy 4).

All experiments (except on real data) can be reproduced by using the R package scoringTools (see Ref. [7] and Section 2 in the Supplemental Material).

References

[1] B. Anderson, *Using Bayesian networks to perform reject inference*, Expert Systems with Applications 137 (2019), pp. 349–356.

[2] J. Banasik and J. Crook, *Reject inference, augmentation, and sample selection*, European Journal of Operational Research 183 (2007), pp. 1582–1594. Available at `http://www.sciencedirect.com/science/article/pii/S0377221706011969`.

[3] G. Celeux and G. Govaert, *A classification em algorithm for clustering and two stochastic versions*, Computational statistics & Data analysis 14 (1992), pp. 315–332.

[4] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-supervised Learning*, Adaptive computation and machine learning, MIT Press, 2010, Available at `https://books.google.fr/books?id=A3ISEAAAQBAJ`.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the em algorithm*, Journal of the Royal Statistical Society: Series B (Methodological) 39 (1977), pp. 1–22.

[6] B. Efron, *The efficiency of logistic regression compared to normal discriminant analysis*, Journal of the American Statistical Association 70 (1975), pp. 892–898.

[7] A. Ehrhardt, *Credit Scoring Tools: the `scoringTools` package* (2020). Available at `https://CRAN.R-project.org/package=scoringTools`, `https://adimajo.github.io/scoringTools`.

[8] A. Feelders, *Credit scoring and reject inference with mixture models*, International Journal of Intelligent Systems in Accounting, Finance & Management 9 (2000), pp. 1–8. Available at `http://www.ingentaconnect.com/content/jws/isaf/2000/00000009/00000001/art00177`.

[9] A. Guizani, B. Souissi, S.B. Ammou, and G. Saporta, *Une comparaison de quatre techniques d'inférence des refusés dans le processus d'octroi de crédit*, in *45 emes Journées de statistique.* 2013. Available at `http://cedric.cnam.fr/fichiers/art_2753.pdf`.

[10] Y. Kang, R. Cui, J. Deng, and N. Jia, *A novel credit scoring framework for auto loan using an imbalanced-learning-based reject inference*, in *2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*. IEEE, 2019, pp. 1–8.

[11] N.M. Kiefer and C.E. Larson, *Specification and informational issues in credit scoring*, Available at SSRN 956628 (2006). Available at `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=956628`.

[12] N. Kozodoi, P. Katsas, S. Lessmann, L. Moreira-Matias, and K. Papakonstantinou, *Shal-*

low Self-Learning for Reject Inference in Credit Scoring, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer, 2019, pp. 516–532.

[13] Z. Li, Y. Tian, K. Li, F. Zhou, and W. Yang, *Reject inference in credit scoring using semi-supervised support vector machines*, Expert Systems with Applications 74 (2017), pp. 105–114.

[14] R. Little and D. Rubin, *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, Wiley, 2019, Available at `https://books.google.fr/books?id=OaiODwAAQBAJ`.

[15] R.A. Mancisidor, M. Kampffmeyer, K. Aas, and R. Jenssen, *Deep generative models for reject inference in credit scoring*, Knowledge-Based Systems (2020), p. 105758.

[16] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley Series in Probability and Statistics, Wiley, 2004, Available at `https://books.google.fr/books?id=c2\_fAox0DQoC`.

[17] G. Molenberghs, C. Beunckens, C. Sotto, and M.G. Kenward, *Every missingness not at random model has a missingness at random counterpart with equal fit*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 7 (2008), pp. 371–388.

[18] H.T. Nguyen, *Reject inference in application scorecards: evidence from France*, Tech. Rep., University of Paris West-Nanterre la Défense, EconomiX, 2016. Available at `http://economix.fr/pdf/dt/2016/WP_EcoX_2016-10.pdf`.

[19] T.J. O'neill, *The general distribution of the error rate of a classification procedure with application to logistic regression discrimination*, Journal of the American Statistical Association 75 (1980), pp. 154–160.

[20] F. Shen, X. Zhao, and G. Kou, *Three-stage reject inference learning framework for credit scoring using unsupervised transfer learning and three-way decision theory*, Decision Support Systems 137 (2020), p. 113366.

[21] R. Sutton and A. Barto, *Reinforcement Learning, second edition: An Introduction*, Adaptive Computation and Machine Learning series, MIT Press, 2018, Available at `https://books.google.fr/books?id=uWV0DwAAQBAJ`.

[22] V. Vandewalle, *Estimation et sélection en classification semi-supervisée*, Theses, Université des Sciences et Technologie de Lille - Lille I, 2009. Available at `https://tel.archives-ouvertes.fr/tel-00447141`.

[23] E. Viennet, F.F. Soulié, and B. Rognier, *Evaluation de techniques de traitement des refusés pour l'octroi de crédit*, arXiv preprint cs/0607048 (2006). Available at `http://arxiv.org/abs/cs/0607048`.

[24] H. White, *Maximum likelihood estimation of misspecified models*, Econometrica 50 (1982), pp. 1–25. Available at `http://www.jstor.org/stable/1912526`.

[25] Y. Xia, X. Yang, and Y. Zhang, *A rejection inference technique based on contrastive pessimistic likelihood estimation for p2p lending*, Electronic Commerce Research and Applications 30 (2018), pp. 111–124.

[26] B. Zadrozny, *Learning and evaluating classifiers under sample selection bias*, in *Proceedings of the twenty-first international conference on Machine learning.* ACM, 2004, p. 114.