EconomiX

http://economix.fr

Reject inference in application scorecards: evidence from France

Document de Travail
Working Paper
2016-10

Ha-Thu Nguyen

UMR 7235

Université de Paris Ouest Nanterre La Défense
(bâtiment G)
200, Avenue de la République
92001 NANTERRE CEDEX

Tél et Fax : 33.(0)1.40.97.59.07
Email : nasam.zaroualete@u-paris10.fr

université Paris Ouest Nanterre La Défense

Research Paper

Reject inference in application scorecards: evidence from France

Ha-Thu Nguyen

EconomiX, Université Paris Ouest-Nanterre la Défense (Paris X), 200 Avenue de la République, Bâtiment G, 92001 Nanterre, France; email: hathunguyen12@gmail.com

February 2016

ABSTRACT

Credit scoring models are commonly developed using only accepted Known Good/Bad (G/B) applications, called KGB model, because we only know the performance of those accepted in the past. Obviously, the KGB model is not indicative of the entire through-the-door population, and reject inference precisely attempts to address the bias by assigning an inferred G/B status to rejected applications. In this paper, we discuss the pros and cons of various reject inference techniques, and pitfalls to avoid when using them. We consider a real dataset of a major French consumer finance bank to assess the effectiveness of the practice of using reject inference. To do that, we rely on the logistic regression framework to model probabilities to become good/bad, and then validate the model performance with and without sample selection bias correction. Our main results can be summarized as follows. First, we show that the best reject inference technique is not necessarily the most complicated one: reweighting and parceling provide more accurate and relevant results than fuzzy augmentation and Heckman's two-stage correction. Second, disregarding rejected applications significantly impacts the forecast accuracy of the scorecard. Third, as the sum of standard errors dramatically reduces when the sample size increases, reject inference turns out to produce an improved representation of the population. Finally, reject inference appears to be an effective way to reduce overfitting in model selection.

Keywords: Reject inference, sample selection, selection bias, logistic regression, reweighting, parceling, fuzzy augmentation, Heckman's two-stage correction.

JEL Classification: C51, C52, C53, G21

1 INTRODUCTION

Nowadays, credit scoring is of paramount importance to evaluate credit risk when banks decide whether or not to approve a credit. Such models are traditionally developed based on the repayment performance of previous applicants in the portfolio, assuming that the sample used for the development represents the overall population (Greene, 1998). In case of application scorecard, this assumption is not fully satisfied, because we only get information on the repayment behavior of those who have been accepted and finally booked for credit. The behavior of those who have been declined is unknown. Thus, by selecting only approved applicants and ignoring rejected ones, the modeling sample is intrinsically biased. If the model is to be used only to predict future bad rates of the approved sample (behavior scorecard), there will be no selection bias issue. However, since the model is typically designed to assess the whole through-the-door population to make new decisions on whether accept or reject an application (application scorecard), the bias turns to be a serious issue (see Chandler and Coffman, 1977, and Avery, 1977 for early discussions on this topic). Eisenbeis (1977), Reichert *et al* (1983), Joanes (1993/4), Hand (1998), Feelders (2000) among others, also report that using a model based solely on accepted applicants often generate misleading results. Reject inference techniques is then a way to address this concern.

The importance of reject inference varies, depending on the current decision process, data, and approval/rejection rates. According to Siddiqi (2006), reject inference does not make much significant difference in the following situations: (i) a very high confidence level with a high approval rate assumes all rejects are equal to bads; (ii) a very low confidence level assumes that decisions are randomly made or using a mistaken adjudication tool. In cases with either low or medium approval rates, reject inference helps in identifying creditworthy applications that are currently rejected.

Previous research has proposed different techniques to reduce selection bias (see Banasik and Crook, 2004 for example). While the performance of these techniques has largely been discussed at a theoretical level, very few studies have been done from an empirical point of view, as most datasets on which reject inference techniques have been tested are incomplete or simulated (Åstebro and Chen, 2012). We fill this gap in the present paper. Indeed, a significant contribution and novelty of our empirical analysis is to use a dataset coming from one of the biggest French consumer finance banks, which contains complete information on both rejected and accepted applications, and focus on the results that could be reached if reject inference would occur. The aim of the paper is to shed some new light on the relevance of reject inference procedures in the development of application credit scoring models.

The rest of the paper is structured as follows. Section 2 discusses the issue of selection bias and different reject inference methods in credit scoring: manual estimation, bureau data-based method, reweighting, parceling, fuzzy augmentation and Heckman's correction method. The pros and cons of each technique are also presented in this section, with a special focus on the Heckman's two-stage estimation method which has been widely used to correct selection bias.

Section 3 is devoted to the empirical part of the paper. Using a real dataset of a leading French consumer finance bank, we evaluate the effectiveness of the practice of using reject inference. Finally, Section 4 concludes the paper with a discussion on the key findings of the study.

2 REVIEW OF REJECT INFERENCE TECHNIQUES

In the sense of Rubin (1976) and Little and Rubin (2002), missing data can be classified in one of three groups: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In the first two cases, the missing data mechanism is ignorable. Indeed, missing at random means that the probability of default, given all the exogenous variables of the model, is the same whether an application is accepted or rejected. On the contrary, if the data is MNAR, the missing data mechanism is non-ignorable because extra information on future default is added by human judgment, and then this will change the probability of default. In this case, sample selection bias happens. Missing data must be included in the model to get proper estimates of the outcome.

In this section, we discuss different applicable methods for non-ignorable missing data.

2.1 Manual estimation

Montrichard (2007) presents manual estimation, known as "expert rules", as one of applicable reject inference methods. It corresponds to the case where the analysts use their own experience to manually simulate the performance of the rejects. For example, they argue that if one variable is greater than a specific limit, over 50% of that group will turn to be "bad" after 12 months. They will then select this segment from the declined dataset and randomly assign 50% of the applicants as "good" and the remaining as "bad". For the remaining rejects, they assume an overall 30% bad rate. The use of the method is not statistically justified and it is consequently likely to be too subjective to be reliable. In fact, if one chooses different rates, the results will be completely different.

2.2 Similar in-house or Bureau data-based method

One of the most commonly used practices in reject inference is to obtain external performance data from credit bureaus on rejected applicants, suggested by Ash and Meester (2002), Siddiqi (2006), Barkova *et al.* (2013), and Etimova *et al.* (2013). Using performance data coming from other creditors over an observation period allows lender to infer how the rejects would have performed if they had been approved. It is assumed that default on other products is equivalent to default on the interested product; the new model is then created with consideration of the behavior of the reject. Thus, using real historical data and giving a more realistic perspective of the market and other clients' behaviors are the core advantages of this method.

However, it also contains several downsides. First, obtaining bureau data is costly and time consuming. It requires extra effort for data preparation and extraction. Second, the quality of the data is uncertain. Third, the main assumption of this approach requires a significant bureau match rate. Finally, the approach cannot completely eliminate bias, since rejects without performance information are probably considered to be non-random.

2.3 Reweighting

Reweighting is a well-used technique that involves weighting accepted applications in such a way as to scale up to the total population (Hsia, 1978; Banasik & Crook, 2004, 2007). The technique uses the ratio between the number of approved accounts in a cluster and the number of declines. Based upon this ratio, the approved accounts will be reweighted such that the number of the weights will equal the total number of applications. As accepted scores are monotonically related to the probability of being accepted, we can replace scores by their probabilities, and consider each individual record in spite of clusters. Each accepted record has a probability to be approved (p_A), and correspondingly a sampling weight of $w = (1/p_A)$. A new model using weighted accepts is then estimated.

Simplicity is a major advantage of this technique. It contains, however, several drawbacks. Banasik and Crook (2004), using their data, demonstrate that reweighting does not really improve the performance of the good-bad model. First, the scope for improving on a model parameterized only on the approvals appears small. Second, reweighting applications within an approval sample and adopting a cut-off point based on those approvals do not seem to perform better than an unweighted estimation. Third, reweighting may undermine the ability to apply good/bad knowledge of the population, without giving any compensating advantage. Reweighting in Parnitske (2005) also appears unsuitable to completely get rid of selection bias. Lacking knowledge of the tendency to become delinquent of the rejects, this method can only lead to improvements by chance. Though, combining reweighting and supplementary data improves the results.

2.4 Extrapolation

Extrapolation involves assigning a default status to the rejects, based on the same model that is fitted to the approvals only, and then re-estimating the model (Ash and Meester, 2002). To simulate outcomes for the rejects, we should follow three steps.

Step 1: Construct a G/B model on the approvals (KGB population) as usual to get the scorecard A.

Step 2: Estimate probabilities to be bad that can be assigned to rejects based on outcomes given by the scorecard A. These estimated bad rates will be then used to simulate outcomes on the rejects, either by Monte Carlo parceling methods or by fuzzy augmentation.

The Monte Carlo parceling method simulates a $0/1$ outcome ($y = 1$ if default/bad, $y = 0$ if non-default/good) for each reject. Next, we generate a random number $r \in [0,1]$ from a

uniform distribution. A reject with score s_A, calculated with scorecard A, is simulated with $y = 1$ if and only if $r < p(y = 1 \mid A, s_A)$. This option is realistic. However, since there is a random factor, the outcome may vary if one changes the random seed generator. Several repeated simulations should be done to get more accurate results.

Fuzzy augmentation, in a different way, considers rejects as being both partially "Good" and partially "Bad" ($0/1$ outcomes with probabilities for each reject). Two records are created to each reject: (i) $y = 1$ with weight $p(y = 1 \mid A, s_A)$, and (ii) $y = 0$ with weight $(1 - p(y = 1 \mid A, s_A))$. The overall weight of each reject is equal to 1. Accepted applications are given by real outcomes with weights equal to 1. This option is more complicated. However, it is more stable and repeatable, thanks to the deterministic nature of the augmentation (Montrichard, 2007).

Step 3: Construct a new model on the approvals and the rejects with their simulated outcomes to get the scorecard B.

These extrapolation methods have been discussed with different points of view. Banasik and Crook (2004) argue that extrapolation tends to remain parameter estimates unchanged, but the good/bad rate it provides appears to be insufficiently appropriate. These methods may be quite arbitrary and result incorrectly or lead to a distortion of the actual default data (Kiefer and Larsen, 2006).

However, Parnitske (2005) shows that extrapolation needs no additional risk and leads to a significant improvement to the model. Zeng and Zhao (2014) also state that fuzzy augmentation technique appears more accurate among others. The authors introduce a rule of thumb, in which one can specify a factor to increase the bad rate of rejected applicants.

2.5 Heckman's two-step correction method

Heckman (1979) discusses the bias that results from using nonrandom selected samples to estimate behavioral relationships as an ordinary specification error or "omitted variables" bias.

The performance of the rejected applicants in credit scoring can then be inferred from Heckman's method (Mok, 2009). The model is composed of two mechanisms, which are the selection mechanism z_i and the outcome mechanism y_i, and are modeled by the latent variables z_i^* and y_i^* for observation i $(i = 1, \dots, N)$ respectively:

$$z_i = \begin{cases} 1 & if\ applicant\ is\ accepted, \quad z_i^* \geq 0 \\ 0 & if\ applicant\ is\ rejected, \quad z_i^* < 0 \end{cases}$$

$$y_i = \begin{cases} 1 & if\ applicant\ is\ good, \quad y_i^* \geq 0 \\ 0 & if\ applicant\ is\ bad, \quad y_i^* < 0 \end{cases}$$

The latent variables z_i^* and y_i^* depend on explanatory variables w_i with parameters γ and β, and random errors u_i and e_i:

$$z_i^* = w_i\gamma + u_i$$

$$y_i^* = w_i\beta + e_i$$

The errors u_i and e_i are assumed to be bivariate normally distributed as follows:

$$\begin{bmatrix} u_i \\ e_i \end{bmatrix} \sim \mathrm{N}(\mu, \Sigma) \qquad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

A selectivity problem arises because the performance indicator y_i is observed only when the applicant is accepted ($z_i = 1$) and the errors u_i and e_i are correlated ($\rho \neq 0$). In such a situation, the usual least squares estimators of β are biased and inconsistent.

Consistent estimators are based on the conditional expectation:

$$\begin{aligned} \mathrm{E}[y_i^* | z_i^* > 0] &= w_i\beta + \mathrm{E}[e_i^* | z_i^* \geq 0] \\ &= w_i\beta + \mathrm{E}[e_i^* | u_i^* \geq -w_i\gamma] \\ &= w_i\beta + \rho H_i \end{aligned}$$

Selection bias is corrected by formulating the conditional expectation with the correlation coefficient ρ and the hazard function H_i, also called the inverse of Mills ratio:

$$H_i = \frac{\phi(-w_i\gamma)}{1 - \Phi(-w_i\gamma)}$$

where $\emptyset$ and Φ denote, respectively, the density and the cumulative distribution function for a standard normal random variable.

The parameters in Heckman's correction method can be estimated in two stages. First, the parameter γ is estimated by probit analysis to obtain an estimate $\widehat{H}_i$ of H_i. The maximum likelihood estimation (MLE) of γ can be calculated by means of probit analysis, where the data is assumed to be complete.

$$\mathrm{E}[z_i^*] = \Phi(w_i\gamma)$$

The corresponding log-likelihood function L is derived as follows:

$$L = \sum_i^N z_i \, ln \, \mathrm{P}(z_i = 1) = \sum_i^N z_i \ln \mathrm{E}(z_i) = \sum_i^N z_i \ln \Phi(w_i\gamma)$$

The MLE of γ is obtained by taking γ that sets the derivative of L to zero. An estimate $\widehat{H}_i$ of H_i is also calculated by applying the inverse of Mills ratio.

Second, the parameters β and ρ are estimated by ordinary least squares (OLS) estimation. The OLS estimates of β and ρ are obtained by minimizing the sum of squared errors (SSE) defined as follows:

$$SSE = \sum_{i=1}^N (y_i - \mathrm{E}[y_i^* | z_i^* \geq 0])^2$$

In case of no sample bias, the errors u_i and e_i are not correlated. The presence of sample bias can be then re-checked by testing the null hypothesis that $\rho = 0$.

The Heckman's two-step correction method has been widely recommended by econometricians (see Poirier, 1980; Meng and Schmidt, 1985; and Boye *et al.*, 1989, for early discussions). Greene (1998) analyzes the impact of Heckman's procedure on selection bias and points out that the coefficients of default change significantly compared to the results obtained without correcting bias. Using a proprietary data set, Banasik *et al.* (2003) find that the Heckman-type selection procedure can get an improvement in terms of performance, but the gain is relatively small. Banasik and Crook (2007) also indicate that using a bivariate probit model to address selection bias can improve model accuracy. A recent study by Bucker *et al.* (2013) shows that the model using Heckman's method yields parameter estimates significantly different, both statistically and economically, from the case where rejects are disregarded.

Other researchers demonstrate that Heckman's bivariate two-stage model does not work well for reject inference. Puhani (2000)'s findings show that the estimators of Heckman's model are inefficient. Since the bivariate model is assumed to be linear with errors which are normally distributed and homoscedastic, the estimations are then not reliable when the assumptions are broken. Åstebro and Chen (2012) also agree with this point of view. The authors believe that in practice, we cannot determine a true model specification and a strong sensitivity makes the usefulness of the approach questionable. Hand and Wu (2007), using simulated data, point out that even if the normality assumption holds, when enough customers are rejected and accepted, or when the original accept/reject decision depends largely on the unobserved variables, correcting selection bias through Heckman's method cannot really help.

2.6 How well do these methods?

Literature raises a question of whether we should use reject inference methods, or in other words, how to validate a method and assess its potential gains. Little empirical research on reject inference has been performed on datasets which are incomplete or simulated. Different studies demonstrate that it is difficult to have a reliable model based on reject inference because the assumptions are often strong and easily violated. However, when the information loss due to selection bias is significant and it cannot consequently be neglected, it is recommended to perform reject inference and test different approaches to find out which one is better to reduce the bias.

Table 1 summarizes pros and cons of each reject inference approach.

Table 1: Comparison of different reject inference methods

	Description	Advantages	Disadvantages
Manual estimation	Build a model on the total population: - Use known performance for the accepts - Use experiences to manually simulate the performance of the rejects	- Easy and quick - May reduce selection bias	- Require solid experience to manually simulate the performance of the rejects - Too subjective to be reliable
Bureau data-based method	Build a model on the total population: - On known performance for accepts - On reference performance for rejects	- Use real historical data - Give a more realistic perspective of the market and other clients' behaviors - May reduce selection bias	- Costly and time-consuming - Require extra effort for data preparation and extraction - Uncertain data quality - Require a significant bureau match rate
Reweighting	- Create an "approve/decline" model on the entire dataset. Each record will have a probability to be approved associated with it (p). - Apply a weight (1/p) to scale up to the total population	- Simplicity - Relatively quick - May reduce selection bias	- Require sufficient historical data, then may not work well on small samples with high reject rates - The assumption $P(bad \mid X, rejected) = P(bad \mid X, accepted)$ is strong and generally unrealistic.
Extrapolation	Assign a default status to the rejects, based on the model built on the accepted applicants only, and then re-estimate the model.	- Finely adjust probabilities to common expectations. - May lead to a significant improvement to the model	- Time-consuming - Complicated - Simulations may lead to incorrect results
Heckman's two-step correction	Consider the likelihood of approval and the likelihood of bad performance associated with the predictive attributes.	- Test bias - Have a quite solid basis in statistical theory	- Complicated - Assume that the bivariate model is linear with errors that are normally distributed and homoscedastic. If the assumptions do not hold, the estimates will not be reliable.

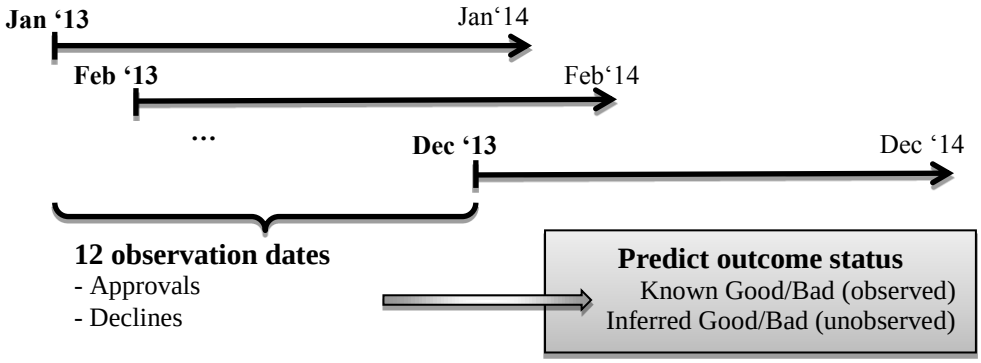
3 REJECT INFERENCE IN PRACTICE

We aim to investigate the performance of reject inference to correct selection bias, using a recent dataset provided by a major French consumer finance bank (hereafter called bank A).

Our sample comprises 198,587 credit histories from January 2013 to December 2013 provided by bank A. For 56,016 applications, the repayment status is known; and all other 142,571 applications are rejected by score (the applications rejected due to a police rule are excluded from the scope of our analysis[1]), then we do not possess any information on their potential repayment. These applications come from bank A's retail loan portfolio for new customers. With such a high-risk portfolio, the rejection rate is consequently high (in excess of 70%); and bias due to selection is typically not ignorable. Thus, reject inference needs to be accounted for to reduce selection bias.

Each of 56,016 applications is observed over a one-year performance window. Figure 1 describes twelve outcome windows selected for scorecard development. Applications are defined as "bad" if they become delinquent for thirty days or more within one year.[2] The worst-ever definition is used. An application is considered as "bad" if the condition holds true at any point over one year, and not at the end of the period. All other applications are defined as "good".

Figure 1: Observations and outcome windows



Next, we create three different datasets (Table 2) for our analysis: (i) "full" dataset contains all observations, both approval and declined applications, (ii) "approve" dataset includes approvals only, and (iii) "decline" dataset includes declines only.

[1] The following rules are applied when selecting rejects to be incorporated into the study:
- If the applications are rejected due to a hard policy rule, which is still in force, then they should not be put into the model.
- The applications rejected due to a soft policy rule and then overridden should be included in the model.
- The applications rejected due to a hard policy rule, which is no longer in force, should also be included.

[2] The definition of bad used in this study is strictly chosen by bank A for this specific portfolio.

Table 2: Full/Approve/Decline datasets

Dataset	#	%
"full"	198,587	100%
"approve"	56,016	28%
"decline	142,571	72%

Our KGB scorecard is developed on the "approve" dataset. As mentioned in Nguyen (2015a), when reviewing variables for possible inclusion in the scorecard, we need to consider the following primary factors: (i) the variables have a significant degree of predictive power (through fine and coarse classing); (ii) they are stable for use; (iii) they have a low correlation to each other; and (iv) they are reasonable enough to explain the business, and also compliant (no legal or ethical restrictions on their use). Only variables that passed this pre-screening step are selected for modeling. Next, we fit a logistic regression model to the dataset. The logistic regression is still the most common method used in credit scoring since it is easy to implement, understand and interpret (Thomas, 2000, and Kocenda & Vojtek, 2006). Moreover, the method works best for binary outcomes, and then it is adopted to build the scorecard since our outcome is typical binary good/bad.

After several trials (see Nguyen, 2015b for criteria allowing selecting the best model among candidates), we find out a scorecard which works best on the "approve" dataset. The final scorecard consists of the following characteristics: *marital status*, *occupation*, *age of the client*, *time at job*, *time at present address*, *household income*, *residential status*, *loan duration*, and *type of credit*. Table 3 on the next pages presents a detailed characteristic analysis.

Table 3 shows that the variables selected in the final scorecard have a good predictive power in overall. The highest predictive ability to separate goods from bads comes from *marital status*, *household income*, *residential status*, *loan duration* and *type of credit*. For example, it is obvious that the risk is lower for married than single or divorced clients, given the fact that *marital status* has an effect on the applicant's responsibility, reliability, and financial wealth.

Table 3: Characteristic analysis

Variable	Attribute	# Good	# Bad	Bad rate	Weight of evidence[3]	Information value[4]
1. Marital status	Married	27 682	873	3.1%	0.48	9.5%
	Divorced, Widowed, Cohabiting	16 314	946	5.5%	-0.13	0.5%
	Single, Separated	9 303	898	8.8%	-0.64	10.0%
		53 299	**2 717**			**20.0%**
2. Occupation	Managers and professionals	6 920	220	3.1%	0.47	2.3%
	Technicians and associated professionals	17 602	701	3.8%	0.25	1.8%
	Service and sales workers	18 502	1 018	5.2%	-0.08	0.2%
	Other jobs	10 275	778	7.0%	-0.40	3.7%
		53 299	**2 717**			**8.0%**
3. Age of the client (in years)	>= 64	10 655	349	3.2%	0.44	3.2%
	50 - 63	16 369	732	4.3%	0.13	0.5%
	40 - 49	13 687	749	5.2%	-0.07	0.1%
	20 - 39	12 588	887	6.6%	-0.32	2.9%
		53 299	**2 717**			**6.7%**
4. Time at jobs (in months)	>= 246	9 019	273	2.9%	0.52	3.6%
	164 - 245 & retired	20 169	808	3.9%	0.24	2.0%
	63 - 163	13 251	824	5.9%	-0.20	1.1%
	<= 62 & unemployed	10 860	812	7.0%	-0.38	3.6%
		53 299	**2 717**			**10.3%**
5. Time at present address (in months)	>= 210	13 572	432	3.1%	0.47	4.5%
	102 - 209	13 399	563	4.0%	0.19	0.9%
	28 - 101	15 840	954	5.7%	-0.17	0.9%
	<= 27	10 488	768	6.8%	-0.36	3.1%
		53 299	**2 717**			**9.4%**
6. Household income (in euros per month)	> 3725	13 657	343	2.5%	0.71	9.2%
	2756 - 3724	13 416	595	4.2%	0.14	0.5%
	2004 - 2755	13 281	721	5.1%	-0.06	0.1%
	<= 2003	12 945	1 058	7.6%	-0.47	6.9%
		53 299	**2 717**			**16.7%**
7. Residential status	House owner	33 114	934	2.7%	0.59	16.4%
	Non-house owner	20 185	1 783	8.1%	-0.55	15.3%
		53 299	**2 717**			**31.7%**

[3] *Weight of evidence* is a measure of how good or bad the accounts are within a particular attribute:

$$Weight\ of\ evidence\ =\ log(\frac{\%good}{\%bad})$$

If the weight is negative, this means there are more bads than goods. If it is positive, this means there are more goods than bads. If it is close to 0, there is a similar number of bads and goods.

[4] The *Information value* measures the predictive ability of a characteristic to separate between good and bad accounts.

$$Information\ value = weight\ \times\ (\%good - \%\ bad)$$

Variable	Attribute	# Good	# Bad	Bad rate	Weight of evidence	Information value
8. Loan duration (months)	<= 35	6 576	183	2.7%	0.61	3.4%
	37 - 48	14 864	466	3.0%	0.49	5.2%
	36	13 220	781	5.6%	-0.15	0.6%
	>= 49	18 639	1 287	6.5%	-0.30	3.8%
		53 299	**2 717**			**13.0%**
9. Type of credit	Installment loans	40 058	1 489	3.6%	0.32	6.4%
	Revolving	13 241	1 228	8.5%	-0.60	12.2%
		53 299	**2 717**			**18.6%**

The Kolmogorov–Smirnov (KS) statistic and the Gini statistic are generally used to measure the discrimination of a scorecard (Nguyen, 2015b). The KS measures the widest spread between cumulative goods and cumulative bads. The divergence between the two curves determines the strength of the scorecard to differentiate good customers from bad ones. The higher the KS, the better the model, since goods are more separated from bads. Like the KS, the Gini coefficient is a quantitative measure of how well the model discriminates between goods and bads but by looking at actual discrimination versus perfect discrimination. A good scorecard in general has a KS greater than 35% and Gini higher than 40%. However, in practice, these thresholds are applicable for scorecards having at least some behavioral variables, e.g. behavioral scorecard or application scorecard for a known-customers portfolio. In our case study, as we have no behavioral variables or high discriminatory variables which can give us a high Gini like other scorecards, we expect then a Gini higher than 35% which could be considered to be satisfactory.

As a result, our scorecard shows a good performance with KS_{dev} of 35% and $Gini_{dev}$ of 46%. The results turn out to be better than expected for such an application scorecard for new customers.

To validate the scorecard, we use a test sample which contains 90,928 credit histories from January to June 2014, which will be observed until June 2015. The choice of the sample is based on the last available repayment behavior information we possess (June 2015). The repayment status is known for 35,726 applications, which are finally financed; and unknown for the remaining 55,202 applications, which are rejected by score. Each of 35,726 applications is observed over a one-year performance window. A "test" dataset which includes these 35,726 approvals and their repayment status is then created for the sake of validation.

Validating the scorecard on the "test" dataset shows a quite significant decrease in terms of performance with KS_{test} of 30% and $Gini_{test}$ of 39%. We use $\Delta Gini$, which is the gap between $Gini_{dev}$ and $Gini_{test}$, as a representative indicator to measure overfitting. In fact, overfitting occurs when the model is typically trained by maximizing its performance on training data, while its efficiency is defined by its ability to perform well on an unobserved data and not by its performance on training data. In this paper, we aim to test different reject

inference techniques to adjust for sample selection bias and find out the one maximizing performance while minimizing overfitting effects.

The following reject inference techniques are incorporated into the study: reweighting, parceling, fuzzy augmentation and Heckman's two-step correction method. The bureau data-based method is not discussed in our empirical analysis since there is no credit bureau in France.[5] The results are summarized in Table 4.

Table 4: Comparison of results of different reject inference methods

	Overall Bad rate	**Known GB Odds/ Inferred GB Odds**	$Gini_{dev}$	$Gini_{test}$	**Overfitting**
KGB model	5%	NA	46%	39%	7%
Reweighting	10%	2.8	44%	38%	6%
Parceling	11%	3.2	38%	39%	-1%
Fuzzy augmentation	43%	19.6	21%	39%	-18%
Heckman's bivariate two-stage model	1%	NS[6]	NS	NS	NS

The overall bad rate in Table 4 is defined as the ratio between the number of bad accounts (both known and inferred bad accounts, but only known bad accounts in case of KGB model) and the total number of good and bad accounts. Since (i) reject inference only provides an estimation of performance if an application had been accepted, and (ii) rejects are more probable to become bad than good, the bad rate with correcting bias should be greater than the one without correcting bias.

In practice, there is a convention that the ratio of the known GB odds to the inferred GB odds should be between two and four times (Siddiqi, 2006). If the ratio is lower, we may infer that the rejects are too similar to the approvals, which is unlikely to be true. If it is higher, then we may be too harsh on the rejects, and may infer too many bad accounts. In this case, the rejects may have too much influence on the final model which is dangerous as their performance has been inferred rather than based on actual history.

Table 4 shows that fuzzy augmentation has a known-to-inferred GB odds ratio of 19.6, given its associated new bad rate is 43%, which is too high. The Heckman's correction method has a non-significant ratio because its associated new bad rate is only 1%. In our case study, the Heckman's method does not produce better results, as expected. Only reweighting and parceling give better results in terms of performance (38% and 39% respectively), while reducing overfitting effect (from 7 points to 6 points and 1 point respectively). The results of KGB model, reweighting and parceling are illustrated in Appendixes A, B and C.

Table 5 displays the estimates of the parameters of the model with and without correcting selection bias, using reweighting and parceling techniques. The results show that reject inference reduces dramatically the sum of standard errors (0.42 vs. 0.99, and 0.58 vs. 0.99).

[5] According to Viola de Azevedo Cunha (2013), the sole public register is the French Central Bank (Banque de France), which is responsible for supplying credit institutions with credit information.
[6] NS stands for non-significant.

Moreover, ignoring rejects may impact the forecast accuracy of the scorecard. In fact, the full model (with reweighting and parceling) yields parameter estimates, which are different, both statistically (when the sign of the estimate is reversed, e.g. *time_at_job2*) and economically (when the estimates change significantly, e.g. *occupation2*, *age1*, *time_at_job4*, *household_income4*, *loan_duration4*), from the case when rejects are disregarded (KGB model).

4 CONCLUSION

Reject inference is of great importance to correct sample selection bias which may affect credit scoring. Literature has mostly focused on presenting different reject inference techniques, and only few efforts have been made to quantify their potential gains. In this paper, we benefit from a real dataset from a known French consumer finance bank to assess the effectiveness of the reject inference techniques. In fact, when applied to a dataset of almost 200,000 new customers requesting credit with a high reject rate in excess of 70 %, our results turn out to be particularly interesting since they have not yet been found by previous researchers. First, we prove that the most suitable reject inference technique is not necessarily the most complicated one. In our case study, reweighting and parceling provide more accurate and relevant results than fuzzy augmentation and Heckman's two-step correction method. Second, we find that ignoring the rejected applicants may impact the forecast accuracy of the scorecard: the full model yields parameter estimates, which are different, both statistically and economically. Third, we demonstrate that reject inference turns out to produce an improved representation of the population since it reduces dramatically the sum of standard errors. Finally, and most importantly, reject inference appears to be an effective way to reduce overfitting in model selection.

On the whole, our results highlight that reject inference can reduce sample bias, and its effect is not modest as it is stated in the literature. Especially in case of a portfolio with a high rejection rate, we cannot ignore rejected applications. Since the techniques are complicated to be implemented, reject inference has been performed with care and caution to find out the most pertinent model developed on the entire through-the-door population.

Table 5: Parameter Estimates

Variables[7]	KGB model			Reweighting			Parceling		
	Coefficient	**SE[8]**	**P-value**	**Coefficient**	**SE**	**P-value**	**Coefficient**	**SE**	**P-value**
Intercept	2.9191	0.0639	<.0001	2.8657	0.0330	<.0001	2.9071	0.0355	<.0001
Marital_status1	0.2487	0.0518	<.0001	0.2746	0.0349	<.0001	0.2450	0.0341	<.0001
Marital_status3	-0.3148	0.0512	<.0001	-0.4020	0.0208	<.0001	-0.2933	0.0218	<.0001
Occupation1	0.4353	0.0769	<.0001	0.2814	0.0236	<.0001	0.4277	0.0234	<.0001
Occupation2	-0.1620	0.0741	0.0289	NS[9]	NS	NS	-0.1588	0.0401	<.0001
Occupation4	-0.2233	0.0515	<.0001	-0.5115	0.0167	<.0001	-0.2377	0.0164	<.0001
Age1	0.2938	0.0797	0.0002	NS	NS	NS	0.3330	0.071	<.0001
Time_at_job1	0.3896	0.0739	<.0001	0.3169	0.0639	<.0001	0.4119	0.067	<.0001
Time_at_job2	0.1379	0.0694	0.047	-0.2193	0.0244	<.0001	0.1142	0.0547	0.0368
Time_at_job4	-0.1837	0.0524	0.0005	NS	NS	NS	-0.2028	0.0257	<.0001
Time_at_present_address1	0.2846	0.0617	<.0001	0.1610	0.0270	<.0001	0.2758	0.0229	<.0001
Time_at_present_address2	0.1718	0.0521	0.001	0.1766	0.0251	<.0001	0.2113	0.0253	<.0001
Household_income1	0.3202	0.0642	<.0001	0.3589	0.0576	<.0001	0.3240	0.0567	<.0001
Household_income4	-0.3634	0.047	<.0001	0.0511	0.0213	0.0164	-0.3527	0.0219	<.0001
Residential_status1	0.7664	0.0449	<.0001	0.7063	0.0382	<.0001	0.7736	0.0385	<.0001
Loan_duration4	-0.4761	0.0413	<.0001	-0.0918	0.0161	<.0001	-0.4753	0.0147	<.0001
Type_of_credit2	-0.6379	0.0411	<.0001	-0.6999	0.0160	<.0001	-0.6456	0.0156	<.0001
Sum of Standard Errors		**0.9971**			**0.4186**			**0.5853**	

[7] The following attributes are not presented in Table 5 :
- The reference attribute in each variable with its weight of evidence in absolute value closest to 0;
- The attributes which are not statistically significant at the 0.05 level;
- The attributes which are correlated with each other.

[8] SE stands for standard error.

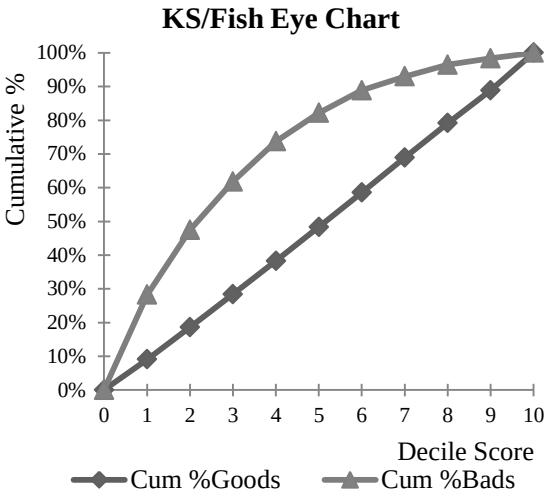
[9] NS means non-significant.

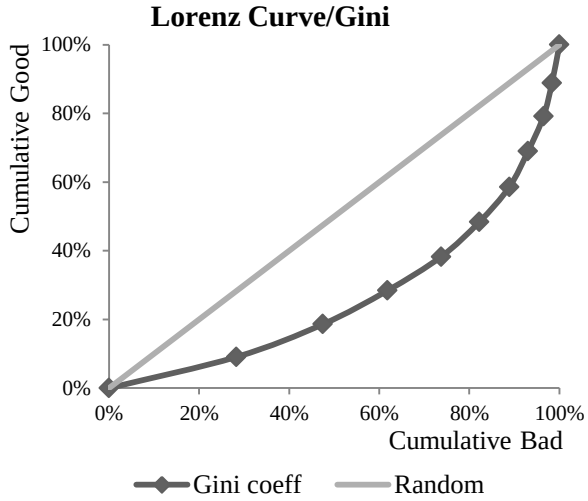
ACKNOWLEDGEMENTS

I am heartily thankful to my supervisor, Professor Valérie Mignon, for her valuable comments on earlier drafts of this paper. I would also like to thank the providers of the data, who prefer to remain anonymous. All errors are my own.

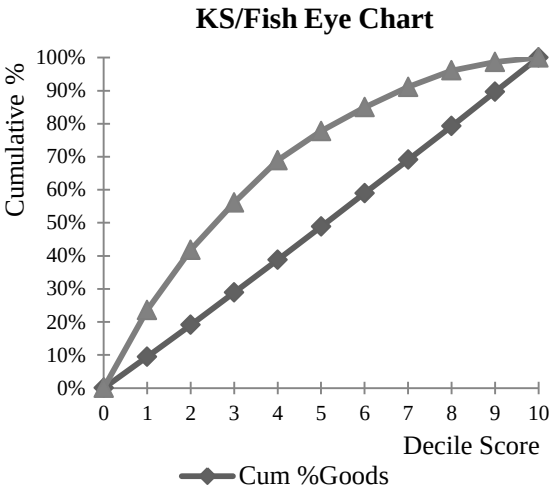
Appendix A: KGB model's performance statistics

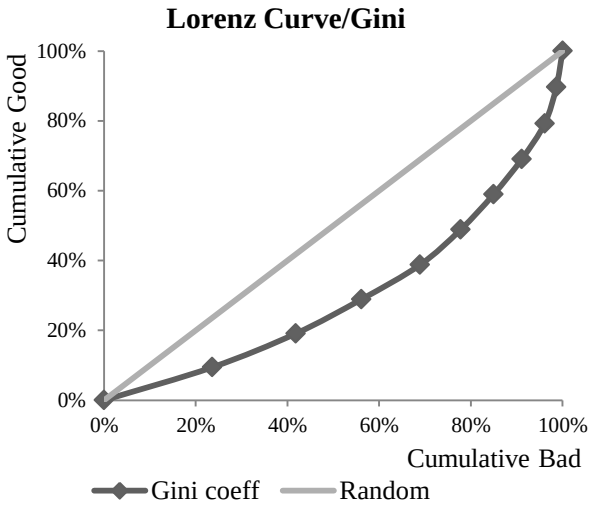
Development population





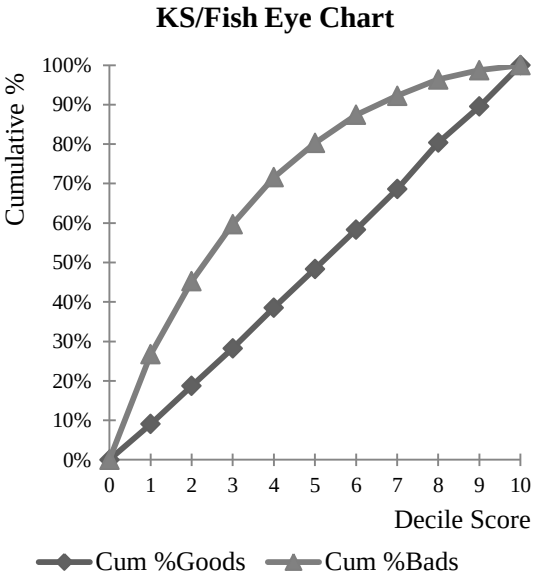
Test population

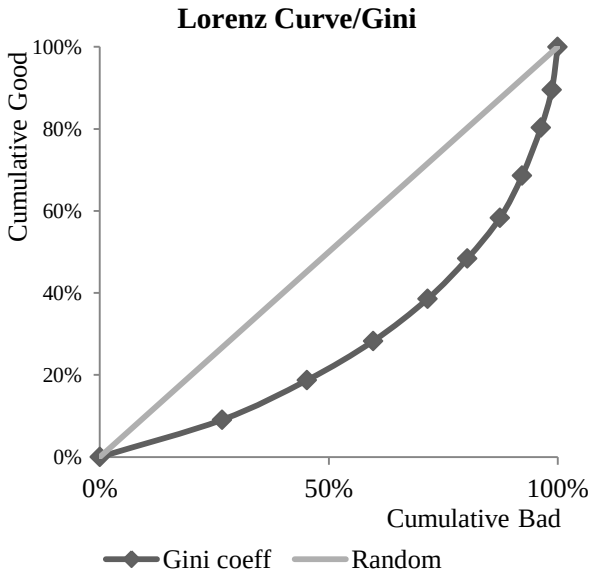




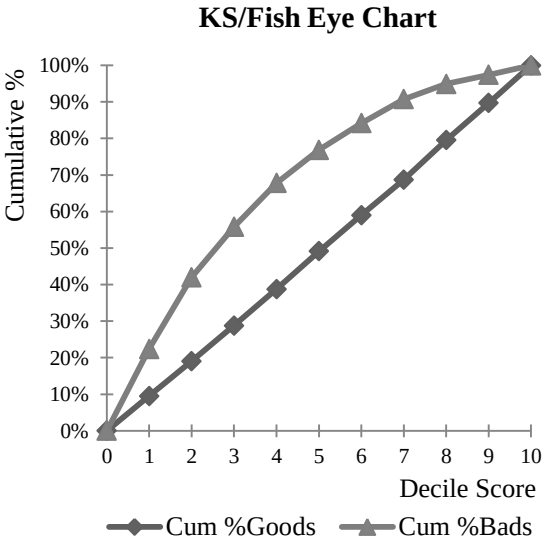
Appendix B: Full model's performance statistics using reweighting technique

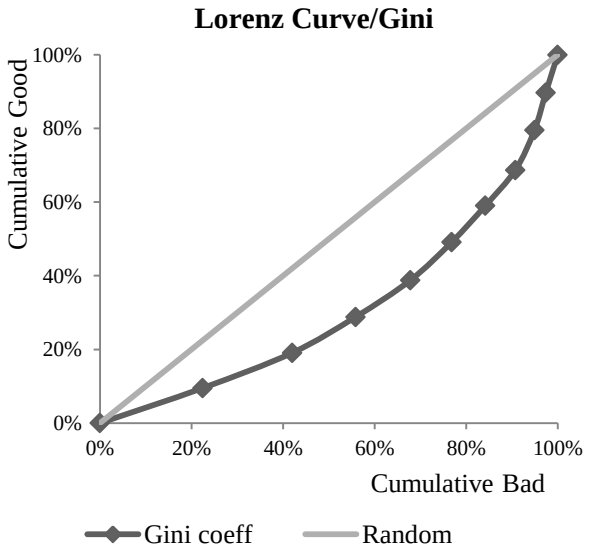
Development population





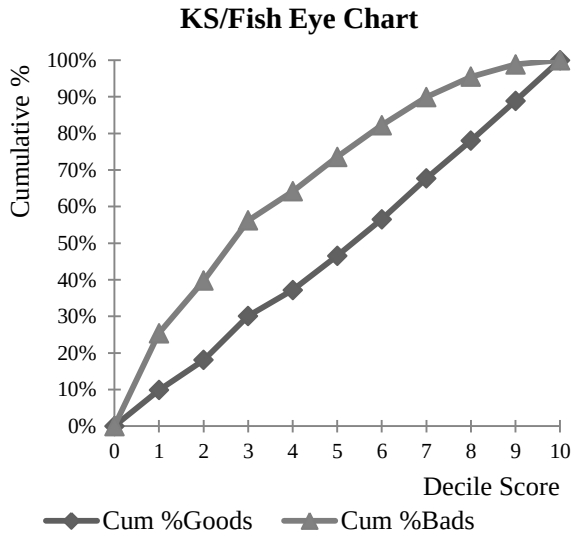
Test population

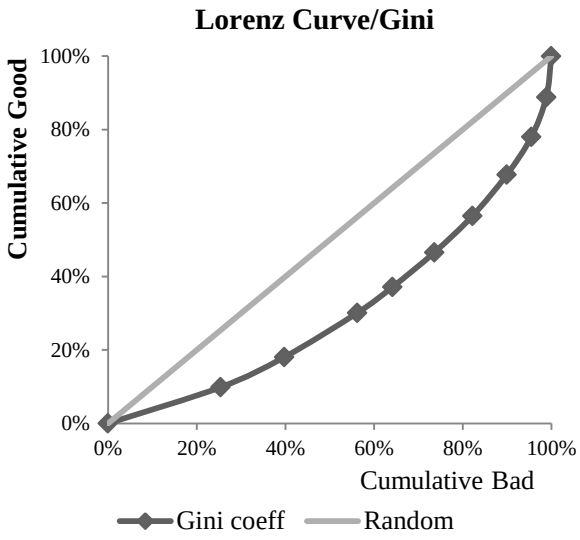




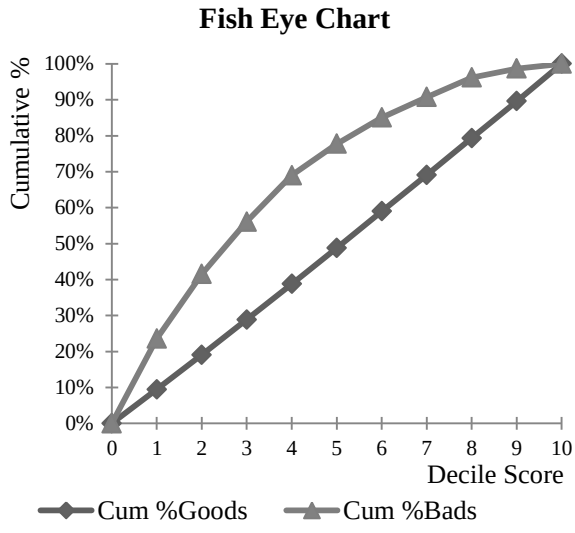
Appendix C: Full model's performance statistics using parceling technique

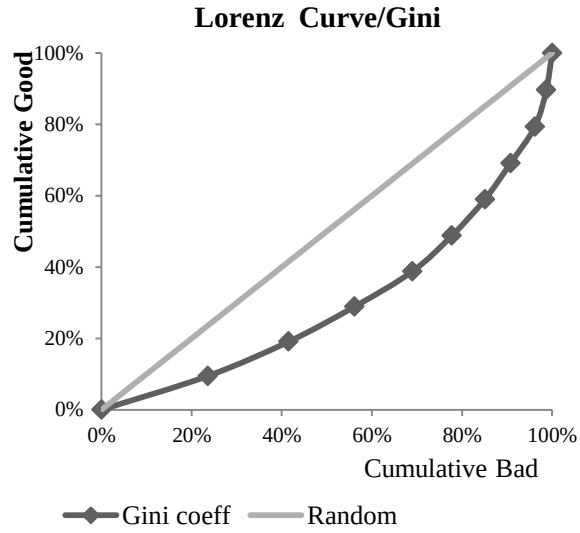
Development population





Test population





REFERENCES

Ash, D. and Meester, S. (2002). Best practices in reject inferencing. Conference on Credit Risk Modeling and Decisioning, Wharton Financial Institutions Center, Philadelphia.

Åstebro, T. and Chen G. (2012). Bound and collapse Bayesian reject inference for credit scoring. Journal of the Operational Research Society **63**, 1374-1387.

Avery, R.B. (1977). Credit scoring models with discriminant analysis and truncated samples. Unpublised paper, Graduate School of Industrial Administration, Carnegie-Mellon University.

Banasik, J. and Crook, J. (2004). Does reject inference really improve the performance of application scoring models?. Journal of Banking and Finance **28**, 857-874.

Banasik, J. and Crook, J. (2007). Reject inference, augmentation, and sample selection. European Journal of Operational Research **183**, 1582-1594.

Banasik, J., Crook, J., and Thomas, L.C. (2003). Sample selection bias in credit scoring models. Journal of the Operational Research Society **54**, 822–832.

Barkova, I., Glennon, D., and Palvia, A. (2013). Sample selection bias in acquisition credit scoring models: an evaluation of the supplemental-data approach. Journal of Credit Risk **9**(3), 77-117.

Boyes, W.J., Hoffman, D.L., and Low, S.A. (1989). An econometric analysis of the bank credit scoring problem. Journal of Econometrics 40(1), 3-14.

Bücker, M., van Kampen, M. and Krämer, W. (2013). Reject inference in consumer credit scoring with nonignorable missing data. Journal of Banking and Finance **37**, 1040-1045.

Chandler, G.G. and Coffman, J.Y. (1977). Using credit scoring to improve the quality of consumer receivables: legal and statistical implications. Paper presented at the Financial Management Association meetings, Seattle, Washington.

Eisenbeis, R.A. (1977). Pitfalls in the application of discriminant analysis in business, finance and economics. Journal of Finance **32**(3), 875-900.

Etimova, R., Gregory, L., and Marcheva, M. (2013). An empirical survey of reject inference techniques. Experian.

Feelders, A.J. (2000). Credit scoring and reject inference with mixture models. International Journal of Intelligent Systems in Accounting, Finance & Management **9**(1-8).

Greene, W. (1998). Sample selection in credit-scoring models. Japan and the World Economy **10**(3), 299-316.

Hand, D. J. (1998). Reject inference in credit operations. In *Credit Risk Modeling: Design and Application,* Mays. E. (eds), pp. 181-190, Amacom.

Hand D.J. and Wu I-D. (2007). Handling selection bias when choosing actions in retail credit applications. European Journal of Operational Research **183**, 1560–1568.

Heckman, J.J. (1979). Sample selection bias as a specification error. Econometrica **47**, 153-161.

Hsia, D.C. (1978). Credit scoring and the equal credit opportunity act. The Hastings Law Journal **30**, 371-448.

Joanes, D.N. (1993/4). Reject inference applied to logistic regression for credit scoring. IMA Journal of Mathematics Applied in Business and Industry **5**, 35-43.

Kiefer, N.M. and Larson C.E. (2006). Specification and informational issues in credit scoring. International Journal of Statistics and Management Systems **1**, 152-178.

Kocenda, E., and Vojtek, M. (2011). Default predictors in retail credit scoring: evidence from Czech banking data. Emerging Markets Finance and Trade **47**(6), 80–98.

Little, R.J.A. and Rubin, D.B. (2002). Statistical analysis with missing Data. Wiley.

Meng, C.-L. and Schmidt, P. (1985). On the cost of partial observability in the bivariate probit model. International Economic Review **26**(1), 71-85.

Mok, J-M. (2009). Reject inference in credit Scoring. BMI Paper. Available from http://www.few.vu.nl/en/Images/werkstuk-mok_tcm244-91398.pdf

Montrichard, D. (2007). Reject inference methodologies in credit risk modeling. Canadian Imperial Bank of Commerce. Available from http://analytics.ncsu.edu/sesug/2008/ST-160.pdf

Nguyen, H.T. (2015a). How is credit scoring used to predict default in China?. Working paper. EconomiX, Université Paris Ouest-Nanterre la Défense (Paris X).

Nguyen, H.T. (2015b). Default predictors in credit scoring: evidence from France's retail banking institution. Journal of Credit Risk **11**(2), 41–66.

Parnitzke, T. (2005). Credit scoring and the sample selection bias. Working Paper on Risk Management and Insurance.

Poirier, D.J. (1980). Partial observability in bivariate probit models. Journal of Econometrics **12**(2), 209–217.

Puhani, P.A. (2000). The Heckman correction for sample selection and its critique. Journal of Economic Surveys **14**(1), 53–68.

Reichert, A.K., Cho, C.C. and Wagner, G.M. (1983). An examination of the conceptual issues involved in developing credit scoring models. Journal of Business and Economic Statistics **1**(2), 101-114.

Rubin, D.B. (1976). Inference and missing data. Biometrika **63**(3), 581-592.

Siddiqi, N. (2006). Credit risk scorecards: developing and implementing intelligent credit scoring. Wiley.

Thomas, L.C. (2000). A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting **16**, 149-172.

Viola de Azevedo Cunha, M. (2013). Market integration through data protection. Law, Governance and Technology **9**.

Zeng, G. and Zhao, Q. (2014). A rule of thumb for reject inference in credit scoring. Mathematical Finance Letters, 2014 (2).