SOLVING SAMPLE SELECTION BIAS IN CREDIT SCORING: THE REJECT INFERENCE

Gabriele Sabato

Royal Bank of Scotland[1]

1. Introduction

Nonrandom samples may present a significant problem in credit scoring. In general, the developer of a credit scoring system possesses only the behavioural information of accepted applicants. However, the scoring model is to be used to evaluate applicants who are drawn, arguable randomly, from the entire population. Assuming that accepted applicants were qualitatively different from individuals whose application were rejected, developing a scoring model on a sample that includes only accepted applicants may introduce sample selection bias and lead to inferior classification results (see Hand (1998) and Greene (1998)). Methods for coping with this problem are known as reject inference techniques.

Some statisticians argue that reject inference can solve the nonrandom sample selection problem (e.g. Copas and Li (1997), Joanes (1994), Donald (1995) and Green (1998)). In particular, reject inference techniques attempt to get additional data for

[1] The material and the opinions presented and expressed in this article are those of the author and do not necessarily reflect views of Royal Bank of Scotland. E-mail address: gabriele.sabato@rbs.com Tel.: +31 6 51 39 99 07. Address: Group Credit Risk, Paasheuvelweg 25, (BT3345), 1105BP Amsterdam, The Netherlands.

rejected applicants or try to infer the missing performance (good/bad) information[2]. The most common methods explored in the literature are: enlargement, reweighting and extrapolation (see Ash and Meester (2002), Banasik et al. (2003), Crook and Banasik (2004) and Parnitzke (2005)). However, some authors (e.g. Hand and Henley (1993)) demonstrate that the reject inference methods typically employed in the industry are often not sound and rest on very tenuous assumptions. They point out that reliable reject inference is impossible and that the only robust approach to reject inference is to accept a sample of rejected applications and observe their behaviour.

In this paper, we analyze the reasons to use reject inference and we assess the different proposed solutions from a statistical and business related point of view. However, in contrast with most of the available literature, we consider the business perspective more relevant than the statistical one in the financial industry context. As such, we conclude that increasing the prediction accuracy of scoring models should not be regarded as the main goal of reject inference techniques. The possibility of including rejects in the development sample should be considered, instead, as an opportunity to replicate the experience and the decision taken by underwriters, credit analysts or branch managers when assessing applicants' creditworthiness.

Aligning a new scoring model to underwriters' risk assessment will help them to better understand the way the model works and takes the accept/reject decision. This will likely facilitate the introduction of an automated decision system for a product that was

[2] Depending on the chosen binary dependent variable, "good" and "bad" will have different meanings. For credit risk models, these terms are usually associated with non-defaulted and defaulted clients, respectively, as observed at least one year after the client has been booked. Following Basel II, the default event is usually defined as 90 days past due on a financial obligation.

previously manually underwritten and will lower the number of requests to override the system decision increasing the efficiency of the acquisition process.

With regards to reject inference methodologies, most of the literature focuses on how to infer the missing performance of the rejected clients without considering the significant value of the accept/reject information. Although the most common approaches to reject inference (e.g. Hand (2002), Ash and Meester (2002) and Crook and Banasik (2004)) are extremely valuable from the statistical point of view, we believe that financial institutions should follow a more practical method when developing their application models in order to guarantee the successful implementation of their systems. We are convinced that scoring models should not be judged only looking at their performance metrics (e.g. discriminatory power, accuracy, stability), but also based on their comprehensibility, simplicity, level of implementation efforts required and level of overrides that would generate[3].

Finally, we propose a practical approach that allows to make use of the rejected applicants when developing a new scoring model. First, we develop a model to predict the probability of default using only accepted clients and we apply it on the entire sample (accepted and rejected clients). Then, we use the reject rate (RR) to "correct" the observed good/bad odds (O-G/B odds) and find out what would have been the rejected good/bad odds (I-G/B odds). Ultimately, we combine the O-G/B odds and the I-G/B odds in order to derive the real good/bad odds (R-G/B odds), similar to the one that we would have observed if rejected clients would have been accepted.

The remainder of the article is structured as follows. In Section 2, we review some of the most relevant research related to reject inference methodologies for credit scoring.

[3] For a more detailed discussion on this topic, see Sabato (forthcoming).

In Section 3, we extensively analyze the proposed methodology from both a theoretical and an empirical point of view. Data from an unsecured personal loans portfolio of a Brazilian bank is used to test the proposed technique. In Section 4, we submit our conclusions.

2. Review of the relevant research literature

2.1 The missing data problem

Credit scoring models are used to risk rank new or existing clients on the basis of the assumption that the future will be similar to the past. If an applicant or an existing client had a certain behavior in the past (e.g. paid back or not his debt), it is likely that a new applicant or client, with a similar risk profile, will show the same behavior. As such, in order to develop a credit scoring model, we need a sample of past applicants or clients' data related to the same product as the one we want to use our scoring model for. If historical data from the bank is available, an empirical model can be developed. When banks do not have data or do not have a sufficient amount of data to develop an empirical model, an expert or generic model is the most popular solution[4].

When a data sample covering the time horizon necessary for the statistical analysis (usually minimum one year) is available, the performance of the accepted applicants can be observed. We define performance as the default or non-default event associated with each client[5]. This binary variable is the dependent variable used to run the regression

[4] Expert scorecards are based on subjective weights assigned by an analyst, while generic scorecards are developed on pooled data from other banks operating in the same market. For a more detailed analysis of the possible solutions that banks can consider when not enough historical data is available, see Sabato (2008).
[5] See note number 1.

analysis. The characteristics of the client at the beginning of the selected period are the predictors.

If we assume some vector of variables $x=(x_1,\ldots, x_k)$ is completely observed for each applicant, based on the information that is filled in on the loan application form and the credit history of the applicant obtained by the central credit bureau, the binary dependent variable y, instead, is observed for accepted applicants, but missing for the rejected ones. We associate the default event with y=1 and the non-default with y=0 and we define an auxiliary variable a, with a=1 if the applicant is accepted and a=0 in case the applicant is rejected. As such, y is observed only if a=1 and missing when a=0.

Following Little and Rubin (1987), we can classify the missing default information into three type of cases:

- Missing completely at random (MCAR), when the probability of observing y does not depend on the value of y, nor on the value of x. This means that the probability of being selected in group a=1 is identical for all cases.
- Missing at random (MAR), when acceptance depends on x, but conditional on x does not depend on y. In this case, the fraction of y=1 for each subgroup a=1 and a=0 should be the same.
- Missing not at Random (MNAR), when the missing of the y depends on x and y. If we do not include the MNAR data in the development sample, we will introduce selection bias in the model.

In general, the missing data issue is said to be *ignorable* if MAR (or MCAR) applies. In the MNAR case, the missing data issue is called *non-ignorable*.

Some authors (e.g. Hand and Henley (1994) and Feelders (1999)) have analyzed and tried to solve the missing data issue for rejected applicants assuming that the selection mechanism was *ignorable.* Unfortunately, we are convinced that this is not the case in the context of credit scoring. Independently from the tool used to select clients (i.e. manual underwriting or a scorecard), it is reasonable to expect a significant difference in the quality of the two subsamples (accepted/rejected applicants). As such, not including rejected clients in the development sample of a new scoring model will generate selection bias of different degrees based on the reject rate experienced in the sample (i.e. the higher the reject rate, the less ignorable will be the selection bias).

2.2 Reject inference studies

The literature about reject inference methodologies is extensive since many authors during the last 20 years have examined several possible realistic alternatives to infer rejected applicants' behaviour. Rosenberg and Gleit (1994) suggest a very simple approach consisting in granting credit to all applicants for a short time period. This would eliminate the issue of inferring the performance of rejected clients (i.e. we would just need to observe it), but would generate significant costs in terms of impairment charges destroying value and increasing the reputational risk for the financial institution[6]. This solution seems to be unrealistic in today's economic environment.

[6] Reputational risk is the potential that negative publicity regarding an institution's business practices, whether true or not, will cause a direct or indirect loss to the institution. With the current economic crisis, this topic has become very relevant for most supervisory authorities that want to ensure that financial institution will not offer anymore credit to clients above the level that they can reasonably afford. These new rules are known in UK as "Treat Customer Fairly" (TCF). For more details, see http://www.fsa.gov.uk/Pages/Doing/Regulated/tcf/index.shtml.

A similar solution is proposed also from Hand (2002), but in a more reasonable version. He suggests a soft accept/reject threshold that would allow to accept some applicants (not all) that would have been otherwise rejected. This approach is also likely to lead to hard-to-justify, additional impairment costs. Crook and Banasik (2004), instead, do not recommend accepting applicants below cut-off, but just assigning a higher weight to cases near the cut-off with the idea that these cases would be a good proxy for the rejected applicants[7]. This method is known as re-weighting (or augmentation) technique.

Several authors have applied various types of extrapolation to infer the performance of rejected applicants (see for example Meester (2000) and Ash and Meester (2002)). Crook and Banasik (2004), in particular, compare the re-weighting and extrapolation methodologies. They find that both methods do not provide significant benefits in terms of improving prediction accuracy on a model developed only on accepted clients, even when a very large proportion of applicants is rejected.

Heckman's (1979) two stage bivariate probit model has also been proposed for reject inference. This approach does not assume that the samples for the accepted and rejected regions are similar. Technically, the loan granting decision and the default model can be described as a two-stage model with partial observability. Other researchers (e.g. Boyes et al. (1989), Greene (1998) and Jacobson and Roszbach (1999)) have used this

[7] Cut-off is the threshold that is generally set during the implementation of a scoring model to automate the acquisition process. Applicants below cut-off should be rejected and the ones above accepted. In reality, an area (known as grey area) where applications are referred to underwriters to be better assessed is always set around the cut-off. The bigger this area, the less efficient the application process is going to be.

method, but they have also pointed out that the underlying assumptions are often violated when applied to the reject inference problem[8].

The academic literature on reject inference is large and we have reported only some of the most important studies[9]. However, we have found this research focusing almost entirely on the reject inference problem from a statistical point of view, not considering the business aspect related to it. We strongly believe that this aspect should prevail when trying to develop a new application model to be used in the acquisition process of a financial institution. A small increase or decrease in the prediction accuracy of the model should not be the objective of a reject inference technique, but it should be regarded just as "side effect".

3. Methodology

In the previous Section, we have analyzed the extensive literature that explores different statistical techniques to be applied to solve the sample selection bias problem. As already mentioned, the purpose of this study is not to recommend one statistical methodology in particular, but to focus on the value of the accept/reject decision in business lending. As such, we propose a practical approach that can be used for reject inference in the credit scoring context without adding too much complexity on the statistical side.

[8] Heckman's model is based on the assumptions that: 1) the granting and default equations are fully specified and 2) it applies to continuous variables.
[9] For a more comprehensive overview of reject inference studies see Chen and Astebro (2001) and Parnitzke (2005).

As already mentioned, we are convinced that reject inference methodologies used in credit scoring should not be chosen with the intent of increasing the prediction accuracy of the model, but to ensure that the model will learn from the lending decisions taken in the past (judgmentally or with the help of a previous model). Only including rejected applicants in the development sample, we can guarantee that the new model will take lending decisions similar to the ones taken in the past and, therefore, will be fully comprehensible for underwriters, reducing referral rates, overrides and process inefficiencies.

Our approach is based on the belief that if an applicant has been rejected, it means that his quality has been considered lower than the one of an accepted applicant. Based on this quite plausible assumption, we can also say that the higher the reject rate per score band the lower has been considered the quality of the rejected applicants in that band compared to the accepted ones.

In order to test our method empirically, we use a sample including unsecured personal loans applicants of a Brazilian bank. In particular, we have 4.940 applicants, but we are able to observe the performance only of the 3.588 accepted ones[10]. Twenty two variables have been collected for accepted and rejected clients during the application process or derived after in the system and are available for our analysis[11].

[10] Applications cover the period from January to March 2007. Performance for accepted applicants is observed one year after. A client is defined as defaulted if he is 90 or more days past due. Otherwise the client is considered to be good (i.e. current).

[11] In order to apply the methodology proposed in this paper, a financial institution needs to have collected applicant's information for accepted and rejected clients. The most common application variables used are socio-demographic information about the applicants (e.g. marital status, residence type, time at current address, type of work, time at current work, flag phone, number of children, installment on income, etc). When a credit bureau is available in the market, the information that can be obtained related to the behaviour of the applicant with other financial institutions is an extremely powerful variable to be used in application models.

We first develop a good/bad model on accepted clients. Using a statistical stepwise variable selection procedure, based on a likelihood-ratio test with the significance level set at 20%, eight variables are selected in the model. We then run a logistic regression and we find an acceptable discriminatory power of the developed model (Gini index of 48%).

Then, we apply this model on the entire sample, including rejected applicants, and we segment the sample in ten homogenous risk bands by score. For each band we can observe the RR and the O-G/B odds as reported in Table 1. If we focus on the lower risk bands, we can observe that most of the rejected applicants concentrate into these bands. This first result demonstrates that the lending criteria were not random and the new model is reflecting the same criteria that were used before by the underwriters.

Now, our task is to include these lending criteria into the new model using the value of the information provided by rejected clients. If we would not do so, the new model would be less respectful of those criteria used in the past and would take decisions not always comprehensible to underwriters, generating inefficiencies in the application process (e.g. increasing the number of referred or overridden applicants and decreasing the trust of underwriters in the model output).

Moreover, it is essential to recognize that the new model cannot assume that the G/B odds of the applicants that will get a score lower than 776, for example, is going to be 3.76. This would be a significant mistake. If we would use this O-G/B odds to set the cut-off or to define the pricing for this product, we would take a wrong and dangerous decision. Ultimately, reject inference is crucial from a business perspective.

Table 1. **Score distribution without reject inference**

This table shows the score distribution after applying the model developed only on accepted applicants. In the first column, the final score is grouped into homogenous risk bands. In the second and third columns, the number of accepted and rejected clients is presented. In the fourth and fifth columns, the accept/reject odds and the reject rate are calculated. In the sixth and seventh columns, the observed number of good and bad applicants, between the accepted ones, is depicted. In the eighth column, the good/bad odds is calculated. In the last two columns, the number and the percentage of clients are respectively shown.

Final Score	# Accepted	# Rejected	A/R Odds	Reject Rate	Obs. # Good	Obs. # Bad	Obs. G/B odds	# App.	% App.
<776	219	272	0.81	0.55	173	46	3.76	491	9.94
823	293	200	1.47	0.41	248	45	5.51	493	9.98
860	301	175	1.72	0.37	277	24	11.54	476	9.64
892	359	152	2.36	0.30	328	31	10.58	511	10.34
914	374	124	3.02	0.25	355	25	14.20	498	10.08
937	370	117	3.16	0.24	347	23	15.09	487	9.86
960	421	87	4.84	0.17	402	19	21.16	508	10.28
985	388	93	4.17	0.19	374	14	26.71	481	9.74
1012	429	65	6.60	0.13	411	12	34.25	494	10.00
>1012	434	67	6.48	0.13	427	7	61.00	501	10.14
TOTAL	3588	1352	2.65	0.27	3342	246	13.59	4940	100.00

As such, we propose a practical methodology to add the value of the reject information in the model and observe the real bad rate (or G/B odds) per band, similar to the one that we would have observed if we would have accepted all applicants. We use the RR per each band (i) to correct the O-G/B odds. In particular, the correction factor (CF) is obtained multiplying the RR and the O-G/B odds per band (1). Then, the CF is deducted from the O-G/B odds to obtain the *inferred* G/B odds (I-G/B odds)(2). The I-G/B odds is used to derive the number of rejected clients that would have been good or bad if accepted (see Table 2).

$$CF_i = RR_i * \text{O-G/B odds}_i \quad (1)$$

$$\text{I-G/B odds}_i = \text{O-G/B odds}_i - CF_i \quad (2)$$

Table 2. **Score distribution with reject inference**

This table shows the score distribution after applying the model developed only on accepted applicants, but inferring the performance of rejected applicants. Using the reject rate and the G/B odds, we are able to infer the performance of the rejected clients per score band. In the first column, the final score is grouped into homogenous risk bands. In the second and third columns, the number of accepted and rejected clients is presented. In the fourth and fifth columns, the accept/reject odds and the reject rate are calculated. In the sixth and seventh columns, the observed number of good and bad applicants, between the accepted ones, is depicted. In the eighth column, the good/bad odds is calculated. In the ninth column, the correction factor (CF) used to derive the I-G/B odds is shown. In the following 3 columns, the I-G/B odds is presented and it is used to calculate the number of inferred good and bad between the rejected applicants. Then, the real number of goods and bad and the R-G/B odds is calculated. In the last two columns, the number and the percentage of clients are respectively shown.

Final Score	# Accepted	# Rejected	A/R Odds	Reject Rate	Obs. # Good	Obs. # Bad	Obs. G/B odds	Corr Factor	I G/B odds	I # Bad	I # Good	R # Good	R # Bad	R G/B odds	# App.	% App.
<776	219	272	0.81	0.55	173	46	3.76	2.08	1.68	162	110	283	208	1.36	491	9.94
823	293	200	1.47	0.41	248	45	5.51	2.24	3.28	61	139	387	106	3.65	493	9.98
860	301	175	1.72	0.37	277	24	11.54	4.24	7.30	24	151	428	48	8.92	476	9.64
892	359	152	2.36	0.30	328	31	10.58	3.15	7.43	20	132	460	51	8.93	511	10.34
914	374	124	3.02	0.25	355	25	14.20	3.54	10.66	12	112	467	37	12.76	498	10.08
937	370	117	3.16	0.24	347	23	15.09	3.62	11.46	10	107	454	33	13.67	487	9.86
960	421	87	4.84	0.17	402	19	21.16	3.62	17.53	5	82	484	24	20.20	508	10.28
985	388	93	4.17	0.19	374	14	26.71	5.17	21.55	4	89	463	18	25.26	481	9.74
1012	429	65	6.60	0.13	411	12	34.25	4.51	29.74	2	63	474	14	33.40	494	10.00
>1012	434	67	6.48	0.13	427	7	61.00	8.16	52.84	1	66	493	8	59.60	501	10.14
TOTAL	3588	1352	2.65	0.27	3342	246	13.59	3.72	9.87	302	1050	4392	548	8.01	4940	100.00

Ultimately, this simple reject inference method will provide a classification (defaulted/non-defaulted) for all rejected applicants allowing us to calculate the real G/B odds per score band (R-G/B odds). This is the G/B odds that minimizes the sample selection bias introduced with the acquisition process and provides a realistic estimate of risk per score band. This is the measure that the financial institution needs to use to set any kind of strategy associated with this portfolio (e.g. cut-off, profitability, etc.). Not applying reject inference is likely to lead to a significant underestimation of risk for each score band and for the lowest ones, in particular.

We are convinced that a new application model should be developed using a sample enhanced with reject inference independently from the level of improvement in the Gini

or in any other measure of prediction power of the model. Eventually, the goal of reject inference is to ensure that the development sample of any scoring model provides a truthful and realistic representation of risk. The choice of developing a scoring model on a sample including selection bias can be acceptable from a statistical point of view (if the bias are proved to be minor), but will always lead to wrong or inefficient solutions from the business perspective.

4. CONCLUSIONS

In this paper, we have addressed the subject of reject inference. Surveying several other studies focusing on a similar theme and adding a new methodology, we compare and expand upon the received evidence and conclusions with several important findings.

First, in contrast with most of the recent literature, we believe that sample selection bias in credit scoring should not be considered only from the statistical point of view. Although reject inference can improve the prediction accuracy of an application model, we have explained that most of benefits are going to be reaped on the business side, improving efficiency and comprehensibility of the application process (e.g. reducing the level of referred and overridden applicants).

Second, we have demonstrated that setting any risk strategy (e.g. setting cut-off or calculating profitability measures) based on a sample including only accepted clients would lead to wrong and dangerous decisions. Rejected applicants must contribute to provide a meaningful and truthful picture of risk and reject inference is the best way to incorporate their valuable information into scoring models.

Last, we have presented a practical methodology that can be used to infer rejected applicants' performance and include them in the development sample. This is not a sophisticated statistical methodology, but provides acceptable and easy-to-understand

results. We have tested this methodology on a sample of Brazilian unsecured personal loans and we have found positive results.

We conclude that reject inference should be regarded with appropriate attention from financial institutions willing to automate their acquisition process in order to ensure that their models will be accepted and understood by the underwriters. Only in this way will banks make sure to reap the highest level of benefits in terms of efficiency and soundness of acquisition strategies. The increase in the prediction accuracy of scoring models should not be considered as the main goal of reject inference techniques. Instead, we are convinced that reject inference should be used to enhance development samples of all application models regardless of the statistical benefits that may, or may not, bring.

References

Ash, D. and S. Meester, (2002), "Best Practices in Reject Inferencing". Presentation at Credit Risk Modelling and Decisioning Conference, Wharton Financial Institutions Center, Philadelphia, May 2002.

Astebro, T. and G. Chen, (2001), "The Economic Value of Reject Inference in Credit Scoring". In L. C. Thomas, J. N. Crook and D. B. Edelman (eds.): "Credit Scoring and Credit Control VII," Proceedings of Conference held at University of Edinburgh.

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

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

Copas, J. B. and H. G. Li, (1997), "Inference for non-random samples (with discussion)", *Journal of the Royal Statistical Society*, Ser. B, Nr. 59, pp. 55-95.

Crook, J. N. and J. L. Banasik, (2004). "Does reject inference really improve the performance of application scoring models?", *Journal of Banking and Finance*, Vol. 28, Nr.4, pp. 857-874.

Donald, S. G., (1995), "Two-step estimation of heteroskedastic sample selection models", *Journal of Econometrics, Vol.* 65, Nr. 2, pp. 347-380.

Feelders, A. J., (1999): "Credit scoring and reject inference with mixture models," *International Journal of Intelligent System in Accounting, Finance and Management*, Vol. 8, Nr.4, pp. 271-279.

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

Hand, D. J. (1998), "Reject inference in credit operations," in Credit Risk Modeling: Design and Application (ed. E. Mays), pp. 181-190, AMACOM.

Hand, D. J. (2002), "Measurement and prediction models in consumer credit". Presentation at Credit Risk Modelling and Decisioning Conference, Wharton Financial Institutions Center, Philadelphia, May 2002.

Hand, D. J. and W.E. Henley, (1993), "Can reject inference ever work?", *IMA Journal of Mathematics Applied in Business and Industry*, Vol. 5, Nr.4, pp. 45– 55.

Hand D.J. and W.E. Henley, (1994), "Inference about rejected cases in discriminant analysis". In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, editors, "New approaches in classification and data analysis", Springer, New York, pp. 292–299.

Heckman, J. J., (1979), "Sample selection bias as a specification error", *Econometrica*, Vol. 47, Nr. 1, pp. 153-161.

Jacobson, T., and K. F. Roszbach, (1999), "Evaluating bank lending policy and consumer credit risk," in *Computational Finance 1999* (edited by Yaser S. Abu-Mostafa et al.) the MIT Press.

Joanes, D. N., (1994), "Reject inference applied to logistic regression for credit scoring," *IMA Journal of Mathematics Applied in Business & Industry*, Vol. 5, Nr.1, pp. 35-43.

Little, R. J. A. and D. B. Rubin, (1987), "Statistical Analysis with Missing Data", Wiley, New York.

Meester S., (2000), "Reject inference for credit scoring model development using extrapolation", Mimeo, CIT Group.

Parnitzke, T., (2005), "Credit Scoring and the Sample Selection Bias", Working paper, www.defaultrisk.com.

Rosenberg, E. and A. Gleit, (1994), "Quantitative methods in credit management: A survey", *Operations Research*, Vol. 42, Nr. 4, pp. 589–613.

Sabato, G. (2008), "Managing credit risk for retail low-default portfolios". In "Credit Risk: Models, Derivatives and Management", N. Wagner Ed., Chapman & Hall/CRC Financial Mathematics Series.

Sabato, G. (forthcoming), "Credit Risk Scoring Models". In "Encyclopedia of Quantitative Finance", Rama Cont Ed., Wiley and Sons, New York.