Reject Inference in Online Purchases

LENNART MUMM

Master's Thesis at Klarna AB
Supervisors: Jonas Adolfsson and Mikael Hussain
Examiner: Tatjana Pavlenko

May, 2012

Acknowledgements

In the process of creating this master's thesis valuable help and encouragement has been provided by several persons. At Klarna I was supported by both my supervisors, Mikael Hussain and Jonas Adolfsson, but also by other colleagues giving insights into the world of credit scoring. At KTH my examiner Tatjana Pavlenko gave help and suggestions throughout the process.

My sincere thanks to all of you.

Abstract

As accurately as possible, creditors wish to determine if a potential debtor will repay the borrowed sum. To achieve this mathematical models known as credit scorecards quantifying the risk of default are used. In this study it is investigated whether the scorecard can be improved by using reject inference and thereby include the characteristics of the rejected population when refining the scorecard. The reject inference method used is parcelling. Logistic regression is used to estimate probability of default based on applicant characteristics. Two models, one with and one without reject inference, are compared using Gini coefficient and estimated profitability. The results yield that, when comparing the two models, the model with reject inference both has a slightly higher Gini coefficient as well as showing an increase in profitability. Thus, this study suggests that reject inference does improve the predictive power of the scorecard, but in order to verify the results additional testing on a larger calibration set is needed.

Sammanfattning

Långivare strävar efter att, så korrekt som möjligt, avgöra huruvida en potentiell gäldenär kommer att återbetala en erhållen kredit. I avsikt att uppnå detta och kvantifiera risken av en kreditförlust nyttjas matematiska modeller betecknade som scorekort. I denna studie undersöks om scorekortet kan förbättras genom reject inference, det vill säga metoden att inkorporera data från nekade kreditansökanden när scorekortet förfinas. Reject inference-metoden som används heter parcelling. Logistisk regression används för att uppskatta sannolikheten av en kreditförlust baserat på den ansökandes karakteristika. Två modeller skapas, en baserad enbart på godkända köp och den andra med data från såväl nekade köp som godkända, där jämförelser görs mellan modellernas Ginikoefficient och uppskattade lönsamhet. Erhållna resultat ger, vid jämförelse av modellerna, att modellen med data från såväl godkända som nekade kunder både uppvisar en något högre Ginikoefficient och en ökad lönsamhet. Resultaten från denna studie indikerar således att reject inference förbättrar scorekortets prediktiva förmåga, men för att verifiera resultaten erfordras ytterligare tester med en större kalibreringsgrupp.

Contents

1 Introduction 1

2 Missing Data and Its Implications for Reject Inference 2

2.1 Three Missing Data Scenarios 2

2.2 How Applications Are Accepted 3

3 Reject Inference Procedure 4

3.1 Parcelling 4

3.2 Score Intervals and Adjustment of the Bad Rate 4

3.3 Variable Requirement and Random Assignment 5

4 Data 6

4.1 Overview 6

4.2 Segmentation of the Training Set 7

5 Statistical Model 8

5.1 Types of Variables 8

5.2 Logistic Regression 8

5.3 Variable Transformation 11

5.4 Weight of Evidence and Information Value 12

5.5 Performance Measures 13

5.5.1 Gini Coefficient and AUC 13

5.5.2 Profitability 14

5.6 Model Selection and Validation 15

6 Results 16

6.1 Histograms 16

6.2 Gini Coefficient, ROC Curves and Profitability 21

7 Analysis 22

7.1 Analysis of Results 22

7.2 Analysis of Method and Assumptions 24

7.3 Possible Extensions 25

8 Conclusion 25

9 Bibliography 27

List of Figures

5.1 Example of cumulative dummy variables for the explanatory variable age. 12
5.2 An example of the ROC curve from which both Gini coefficient and AUC is calculated. 14
6.1 Histograms of the training set (TS) data scored by the reject inference model. 17
6.2 Histograms of the validation set (VS) data scored by the reject inference model. 18
6.3 Histograms of the calibration set (CS) data scored by the reject inference model. 19
6.4 Histograms of the training set data scored by the reject inference model. The upper plot shows the accepted applications and the lower plot the rejected applications. 20
6.5 A plot of the ROC curves for the model with reject inference and for the one without reject inference. The straight line symbolises a model that assigns the good/bad label completely at random. Data is from the training set. 23
6.6 A plot of the ROC curves for the model with reject inference and for the one without reject inference. The straight line symbolises a model that assigns the good/bad label completely at random. Data is from the validation set. 23
6.7 A plot of the ROC curves for the model with reject inference and for the one without reject inference. The straight line symbolises a model that assigns the good/bad label completely at random. Data is from the calibration set. 24

List of Tables

2.1 Missing outcome performance framework. 2

3.1 The scores in this table are assigned by the existing scorecard, not from the derived model; this is an overview of the input data. The columns are the score, number of bad purchases (B), number of good purchases (G), number of approved purchases (A), probability of a purchase being bad (P_B), adjusted bad rate ($P_{B,adj}$), probability of a purchase being accepted (P_A), and the number of rejected purchases (R). 5

4.1 The number of observations in each data set together with its time window. The letters A and R designate the number of accepted and rejected, respectively. 7

5.1 Variable classification. 8

5.2 A table depicting the possible outcomes of hypothesis testing. The null hypothesis may be anything, but in credit risk context in this study the null hypothesis is that the application is bad. 13

6.1 Gini coefficients for the models, with and without reject inference, for training set (TS), validation set (VS) and calibration set (CS). Note that for the reject inference model the Gini coefficient varies from run to run, which is why the mean values are shown indexed by a μ. Furthermore, the standard deviation indicated by s is added to the table. The Gini coefficient values of the training set are of less interest because of overfitting issues. 21

6.2 The profitability of the models with and without reject inference is measured in terms of estimated loss ratio, R_L, for the validation set (VS) and the calibration set (CS). The values shown are the mean and standard deviation of ten runs. 21

7.1 A summary of how many stores the training and calibration set consist of, respectively. 25

1 Introduction

When a person applies for credit the creditor requires a method for determining whether the request is to be approved or rejected. Mathematical models known as *credit scorecards* (henceforth only denoted as scorecards) are tools that estimate the probability of how the potential debtor will behave if the sought after credit is granted. The scorecards need input variables, i.e. the personal data, e.g. age and income, associated with this particular applicant in order to decide the outcome of the petition, but the exact implementation of this procedure is different for all creditors.

In those instances when credit is granted the applicant's repayment behaviour will be observed by the creditor, and it can be established if the credit was good (the applicant paid the money back) or bad (a default by the applicant). Evidently, repayment behaviour cannot be documented when the credit request is denied, which creates an inherently biased representation of what characteristics lead to a good credit. This can constitute a problem if new rules for credit risk assessment in the updated scorecard are constructed solely from data of previously accepted applicants. Unintentionally, this way one may screen perfectly good applicants merely because the initial decision model rejected them and the ensuing refinements only utilise data from those approved.

Should the model be applied exclusively on that part of the population featuring approved characteristics, this would not be an issue. However, since generally the model is meant to be used on the whole population, neglecting this bias may result in unwanted rejections. In order to devise a scorecard with an efficiency as high as possible creditors would prefer to include the data from the spurned population; the technique known as *reject inference* enables this.

Reject inference is the procedure when the outcome of earlier rejected applicants is modelled in order to be able to label them as good or bad. This way, the creditor is able to improve the existing scorecard without the aforementioned bias.

The aim of this thesis is to investigate the possibility for reject inference on the data of aspiring purchasers of goods from various online stores that use the service of the company Klarna. This company's service is to enable the customer to purchase the desired goods from the e-store in question, get it sent home, and first then pay for it; the whole credit risk is thus transferred from the store to the credit providing company. Therefore it lies in the company's interest to, as accurately as possible, be able to model the credit worthiness of these would-be consumers of goods.

It is to be noted that the above stated problem not exclusively pertains to credit scoring, but exists in various other fields where a selection in some way is performed and further observations are impossible unless the object in question was included in the sample.

2 Missing Data and Its Implications for Reject Inference

2.1 Three Missing Data Scenarios

The need for reject inference arises because of missing outcome performance, i.e. the creditor is unable to observe if rejected applicants are good or bad. Three commonly used missing data scenarios, initially derived by Little and Rubin in [7], are now introduced. The following description is based on the missing data scenario summary in [4, pp. 2-5].

For each applicant a vector of variables $\boldsymbol{x} = (x_1, \ldots, x_k)$ is observed, where this data may stem from the applicant or from other sources. A variable $a \in \{0, 1\}$ is assigned to each applicant, where $a = 0$ denotes an accepted credit and $a = 1$ a rejected credit. For accepted applicants an additional variable $y \in \{0, 1\}$ is introduced, which is missing for rejected ditto. Here, $y = 1$ indicates a bad credit and $y = 0$ a good credit. The three types of missing data scenarios will be described below. The variable y is set to be

- *missing completely at random* (MCAR) if $P(a = 0) = P(a = 0 \,|\, \boldsymbol{x}, y)$, i.e. acceptance is independent of both the data and outcome performance;
- *missing at random* (MAR) if $P(a = 0) \neq P(a = 0 \,|\, \boldsymbol{x}) = P(a = 0 \,|\, \boldsymbol{x}, y)$, i.e. acceptance is dependent on the data but independent of outcome performance;
- *missing not at random* (MNAR) if $P(a = 0) \neq P(a = 0 \,|\, \boldsymbol{x}) \neq P(a = 0 \,|\, \boldsymbol{x}, y)$, i.e. acceptance is dependent on both the data and outcome performance.

Table 2.1: Missing outcome performance framework.

MCAR is applicable when choices were made totally randomly, e.g. by tossing a (perfect) coin. This is, for understandable reasons, not widely used in practice, but may be used initially to get a first data sample to work with.

For the other two scenarios, selection criteria are based on $\boldsymbol{x}$. First out is the case MAR, which is common in practice and occurs when a selection model is fully automated, such that y is observed only when some function of the x_is, $i = 1, \ldots, k$ exceeds a specified threshold or cut-off value. From the MAR assumption an important property follows:

$$P(y = 0 \,|\, \boldsymbol{x}, a = 0) = P(y = 0 \,|\, \boldsymbol{x}, a = 1) = P(y = 0 \,|\, \boldsymbol{x}), \tag{2.1}$$

i.e. the distribution of the observed y is the same as the distribution of the missing y at any fixed value $\boldsymbol{x}$.

A third possibility is the MNAR scenario—the most complicated case—where acceptance is influenced by extraneous factors not recorded in $\boldsymbol{x}$, e.g. an underwriter overriding the decision of the system or an initially declined customer swaying the outcome by perseverance. In this case

$$P(y = 0 \,|\, \boldsymbol{x}, a = 0) \neq P(y = 0 \,|\, \boldsymbol{x}, a = 1), \tag{2.2}$$

meaning that the distribution of the observed y differs from the distribution of the missing y at any $\boldsymbol{x}$ where an extraneous factor influenced the decision.

In cases MCAR and MAR, data is labelled as being *ignorably missing*, in effect meaning that analysis can be performed only on observed performance. On the other hand, for the MNAR cases, data is said to be *non-ignorably missing*, i.e. there is selection bias and the mechanism behind the missing data ought to be included in the model to get good results.

2.2 How Applications Are Accepted

It is pertinent to provide a brief description of the mechanisms behind the procedure determining if the credit applications used as data in this report are to be accepted or rejected. As soon as an order is placed by a customer the provided data is analysed automatically by different algorithms and subject to certain policies. Prior to any calculations applications failing to meet certain minimum requirements, e.g. that customer age has to be 18 or higher, are rejected. If the order is found not to be a policy reject its probability of default is calculated and transformed to a score, which determines if the application is accepted. So far everything is automated and the MAR scenario applies. However, now another policy engine starts, the fraud policy, and if the application is flagged by it it will appear on the manual surveillance list for further inspection. If the decision agent regards the application as dubious he or she can override the scorecard's verdict to approve and instead reject the order.

The two different types of overrides, *low side overrides* and *high side overrides*, are explained in [12]. The former applies when the creditor grants credit to the applicant despite the score falling into a range generally not acceptable. The latter is the converse, i.e. when credit is denied in spite of the score being acceptable.

In regard of these definitions it is clear that the overrides fall on the high side. These overrides could be a reason for concern but, following Hand and Henley in [5, p. 526], as long as the relevant applicant population is defined exclusively of those eliminated by a high side override, in general they will not lead to biased samples. In this study, the number of manually rejected purchases is but a minute fraction of the total amount. From this follows that the MAR scenario seems to fit quite well, and a reject inference procedure applicable for this scenario will be used.

3 Reject Inference Procedure

3.1 Parcelling

In [10] several reject inference procedures are described, and in this test the focus is on *parcelling*, which [8, p. 6] categorises as an extrapolation technique. The reasoning behind the choice of parcelling over another method is that it is quite intuitive and its implementation is not too elaborate. Furthermore, since the aim of this study is to provide an answer to the question if it is advantageous to include reject inference in a future refinement of the model, the inherent randomness of the parcelling algorithm will give an inkling of how much better or worse the new model can get when reject inference is applied. A description of the parcelling algorithm follows, where table 3.1 provides the data used in this study and, additionally, serves as a visualisation of the algorithm in question.

First, there has to be an existing scorecard rejecting or accepting applicants. The response variable of the scorecard is a score, and this score is then further categorised into several intervals, where the number of intervals are decided by the analyst. See section 3.2 for how the scores are allocated in this study. When this is done historical applicants from a specified time window are looked at and then assigned to their corresponding score intervals. For all accepted cases the good or bad outcome is known, whereas the outcome is missing for the rejected applicants. The probability of the accepted applicant being good is given as $P_G = G/A$, with G denoting the number of good orders and A the number of accepted. In the same way $P_B = B/A$, with B as the number of bad orders, is the probability of the applicant being bad. Another quantity of interest is the approval rate $P_A = A/(A+R)$, where R is the number of rejects.

After this has been done, all rejects are scored with the existing model and assigned their corresponding expected P_G and P_B for each interval. Within each interval i $P_G R_i$ of the rejects are labelled as good and $P_B R_i$ as bad. The assignment in each score interval is random. The analyst here has the possibility of adjusting P_B in order to account for a possible difference in bad rate amongst the rejects compared to the accepted. See eq. (3.1). Given that the existing scorecard works as it should it seems reasonable to assume that the bad rate amongst the rejected is somewhat higher.

The third step is to incorporate the rejects into the accepts, where P_G and P_B decide the probability of each reject to be labelled as good or bad, respectively. When all this is done the data set comprises both accepts and rejects, and the response variable of all observations has an entry.

3.2 Score Intervals and Adjustment of the Bad Rate

As mentioned in section 3.1, an explicit number of intervals have to be set for the reject inference procedure. With the scorecard in this study the score the applications receive ranges from 400 to 800, but scores of either extreme are rare. Because

Score	B	G	A	P_B	$P_{B,adj}$	P_A	R
400-500	285	371	656	0.4345	0.4422	0.0914	6,524
501-540	507	3,141	3,648	0.1390	0.1484	0.4035	5,394
541-580	784	21,970	22,754	0.0345	0.0506	0.8544	3,877
581-620	660	41,599	42,259	0.0156	0.0300	0.9628	1,633
621-660	119	45,593	45,712	0.0026	0.0026	0.9920	369
661-700	20	34,323	34,343	0.0006	0.0006	0.9957	150
701-800	1	4,772	4,773	0.0002	0.0002	0.9973	13

Table 3.1: The scores in this table are assigned by the existing scorecard, not from the derived model; this is an overview of the input data. The columns are the score, number of bad purchases (B), number of good purchases (G), number of approved purchases (A), probability of a purchase being bad (P_B), adjusted bad rate ($P_{B,adj}$), probability of a purchase being accepted (P_A), and the number of rejected purchases (R).

no interval should end up with zero or just a handful of orders, all below or equal to 500 are set to fall into the first interval, from then on each interval is set to have a score range of 40 until 700 is reached, and the last one contains all above 700.

Furthermore, a bad rate needs to be set for the rejects; either the same as for the accepted purchases or higher as a conservative measure. In this study the existing bad rate in each segment is adjusted by the known bad rate from a *calibration set* of orders that were accepted regardless of scorecard verdict according to the formula

$$P^{(i)}_{B,adj} = \frac{m_i P^{(i)} + \alpha n_i P^{(i)}_{oot}}{m_i + \alpha n_i}, \quad i = 1, \ldots, k, \quad k \in \mathbb{N}, \tag{3.1}$$

where $P^{(i)}_{B,adj}$ is the adjusted bad rate, $P^{(i)}$ the bad rate of the accepted, m_i the number of accepted orders in the segment, $P^{(i)}_{oot}$ the bad rate in the calibration set, n_i the number of purchases in the out of time sample, α a scaling factor, and k the number of intervals. The basis for this adjustment is found in the assumption that the bad rate of the calibration set is believed to more accurately capture the bad rate amongst the rejects than the bad rate of the accepted purchases. The scaling factor is set to $\alpha = m_i/n_i$ in order to give the orders from the calibration the same weight as those from the training set. If $n_i = 0$ or a value just above zero for interval i, then $P^{(i)}_{B,adj} = P^{(i)}$, explaining the equality between the columns for the highest score ranges.

3.3 Variable Requirement and Random Assignment

Crook and Banasik warn about a potential pitfall in [2, pp. 4-5]: Care has to be taken that the *explanatory variables* of the old scorecard are a subset of the explanatory variables considered as input for the new model, since otherwise data will fall into the category MNAR rendering an omitted variable bias in the estimated parameters. This just mentioned requirement is satisfied in the setting of this study.

An intrinsic property of the parcelling technique, stressed by Montrichard in [8, p. 7], is its random approach to label the rejects as good or bad, in effect meaning that the training data will change for new runs of the algorithm. See section 6 and section 7 for a more thorough discussion about the implications of the random assignment.

4 Data

4.1 Overview

As data set online purchases in Finnish stores in 2011 are used. The orders do not stem from the whole year, but an exact time window will not be specified because of confidentiality. Purchases with dates from the later two thirds of the time window are used as *training set* for both models, i.e. with and without reject inference. These data are what is used to train the models. The calibration set comprises purchases from the whole time window that were accepted regardless of scorecard outcome, where the calibration and training set are disjoint even though they stem from overlapping time periods. With the available data the orders in the calibration set make up a set as close as possible to rejected purchases that still were accepted. Hence, the calibration set enables a way to verify if the model derived with reject inference is superior in terms of predicting applications that would not generally have been accepted by the scorecard. Lastly, there is an *out of time validation set* consisting of the orders from the first third of the time window. See section 5.6 for further details about the validation process.

Around two sevenths of the total amount of purchase attempts are rejected. However, a relatively large proportion of these rejects are not viable to enter into the reject inference procedure, since it is common that a customer with a rejected purchase within seconds or minutes tries to place the same order again—sometimes with the exact same provided details, but oftentimes with the details slightly altered. Hence, in order to avoid multiple counting, renewed attempts of the same purchase are discarded. The removal of these rejections is done with a technique set up for this purpose. Whenever the contact details of a rejected purchase attempt are too similar, or identical, to a previous rejection within a certain time limit, the new rejection is not considered viable for this study since there is no new information contained within.

There are also quite a few of the accepted purchases that do not qualify. What they all have in common is that none of them fall into either of the categories good or bad, instead they are *indeterminate*, i.e. no outcome can be observed. The reasons vary from case to case; two possibilities are that for some reason the store never shipped the goods, or the customer claims that the goods never arrived and thus refuses to pay. All accepted purchases where neither a bad nor a good outcome can be observed are disregarded. Table 4.1 shows a summary of how many observations there are in each of the above mentioned data sets after the removal of duplicate rejections and indeterminate accepted.

Data Set	Obs	*A*	*R*
Training Set	172,105	154,145	17,960
Validation Set	69,404	69,404	0
Calibration Set	1,218	N/A	N/A

Table 4.1: The number of observations in each data set. The letters *A* and *R* designate the number of accepted and rejected, respectively.

4.2 Segmentation of the Training Set

A clear definition is needed to separate good purchases from bad purchases. Conceptually this is straightforward, but in reality some customers pay a very long time after their due date and first after several reminders and debt collection. In this report a purchase is labelled as bad if it has not been paid within 90 days after due date.

An overview of the purchase distribution over the intervals of the score range in the training set is shown in table 3.1. Based on this data the two models, one with reject inference and one without, are derived.

The data presented in table 3.1 show that rejects are more common in the lower score ranges and that their number decreases successively for higher scores—as it ought to be. At first it may seem surprising that there are any rejects at all in the highest score range, but this can be explained by there being an upper amount limit per purchase and that not too many separate purchases may be placed by the same customer before some money is paid. The acceptance rate increases when the score increases, whereby in the lowest range only around one in ten is approved whereas in the three highest ranges there are almost no rejections. The bad rate and its adjusted value are not completely alike in all instances, especially in the score ranges 541-580 and 581-620 there are some differences indicating that the calibration set has a higher proportion of bad applications in these score ranges compared to the training set. The column for the number of accepted shows that the mean score does not separate two equally big halves of the applicant population, but rather that the mean score is skewed to the higher score range.

5 Statistical Model

5.1 Types of Variables

There are different kinds of variables for describing the observations made on the subjects or objects in the study. A summary of existing naming conventions is compiled in [3]. In this text the term *response variable* is used for measurements that are free to vary in response to the explanatory variables where, here, the former is the good/bad label assigned to paying or defaulting customers, and the latter comprises the characteristics of the applicants. Both response and explanatory variables vary in type; table 5.1 lists the possibilities.

- *Nominal variables*, e.g. yes, no; tundra, desert, rainforest. If there are only two possible values the variable is *dichotomous*, otherwise *polychotomous*.
- *Ordinal variables* when there exists a natural ordering between the categories of the variable, e.g. freezing, chilly, warm, hot. Both nominal and ordinal can be referred to as *categorical* variables.
- *Continuous variables* when the observations, in theory, stem from a continuum, e.g. age or time.

Table 5.1: Variable classification.

Additionally, explanatory variables, of any aforementioned type, can be *confounding* or *interacting*. In [6, pp. 55-58] Kleinbaum explains that the former is when a third variable distorts the relation between two variables due to a distinct connection with the two other variables, and the latter when the concurrent influence of two variables on a third is not additive. When devising the model in this report no analysis of confounding or interaction variables are included.

5.2 Logistic Regression

Many different credit scoring techniques exist, and this study will focus on logistic regression since the response variable is dichotomous. Furthermore, as reasoned in [11, p. 6], logistic regression has been shown to work equally well as other, more elaborate, procedures, and it has often been used successfully in the past. Because of the dichotomous response variable ordinary linear regression is unsuitable for various reasons. One reason is that if linear regression is used, the predicted values can become greater than one and less than zero, values theoretically inadmissible, which cannot happen with logistic regression. Another is due to the assumption of *homoscedasticity*, constant variance of the response variable, in linear regression.

In the case of a dichotomous response variable, the variance decreases the higher number of observations have one and the same outcome.

The situation where the aim is to investigate in which way explanatory variables influence the outcome of a dichotomous response variable arises in numerous applications, e.g. in epidemiologic research and in credit scoring, where the latter example is of primary interest in this text. Generally, the explanatory variables can be labelled as $X_1, \ldots, X_p$, $p \in \mathbb{N}$, and in the case when $p > 1$ the just described situation is a multivariable problem and a mathematical model is needed to describe the often complex interrelationships amongst the variables. Many such models exist, e.g. *logistic regression*, *artificial neural networks*, *decision trees*, *discriminant analysis*, etc., but in this text the focus will be on logistic regression. An outline of the model, based on [3] and [6], follows.

The *logistic model* is based on the *logistic function*

$$f(z) = \frac{1}{1 + e^{-z}}, \quad \begin{cases} \lim_{z \to -\infty} f(z) &= 0, \\ \lim_{z \to \infty} f(z) &= 1. \end{cases} \tag{5.1}$$

Considering the limits in eq. (5.1) together with it being a continuous function, it follows that $f(z) : \mathbb{R} \to [0, 1]$. From the logistic function the logistic model is derived. Let $z = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$, $p \in \mathbb{N}$, and consider the dichotomous response variable R (in credit risk context: $R = 0$ denotes a good credit and $R = 1$ a bad credit). The modelled probability can be expressed as a conditional probability, and if it equals the logistic function, the model is defined as logistic, i.e. if

$$P(R = 1 \mid X_1, \ldots, X_p) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{p} \beta_i X_i)}}, \quad p \in \mathbb{N}, \tag{5.2}$$

where the β_is, $i = 0, \ldots, p$, act are the unknown parameters that are to be estimated from the data. With the aim to introduce less cumbersome notation, the conditional probability in eq. (5.2) is henceforth simplified as $P(\boldsymbol{X}) := P(R = 1 \mid X_1, \ldots, X_p)$. Often the logistic model P($\boldsymbol{X}$) is presented in an alternative form called the *logit*, which merely is the simple transformation

$$\text{logit}\, P(\boldsymbol{X}) = \ln\left(\frac{P(\boldsymbol{X})}{1 - P(\boldsymbol{X})}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i. \tag{5.3}$$

Hence, the logit simplifies to a linear sum.

Another point of interest regarding this representation is the ratio $P(\boldsymbol{X})/(1 - P(\boldsymbol{X}))$, since this gives the *odds* for the response variable for a subject with explanatory variables specified by $\boldsymbol{X}$. It follows that the logit is the *log odds*. With this terminology, an interpretation of the intercept β_0 can be derived: β_0 is the log odds when there are no X_i, $i = 1, \ldots, p$; the background log odds. This value could be utilised as a starting point when comparing different odds for a varying number of X_i, $i = 1, \ldots, p$.

Generally speaking, whenever there is a mathematical function g, not necessarily linear, relating the expected value of the independent response variables Y_k

to a linear function of the explanatory variables $X_1, \dots, X_p$, $p \in \mathbb{N}$ of the form

$$g[\mathrm{E}(Y_k)] = \beta_0 + \boldsymbol{X_k^T\beta}, \quad k = 1, \dots, n, \quad n \in \mathbb{N}, \tag{5.4}$$

and where Y_k belongs to the *exponential family of distributions* the model is said to be a *generalised linear model*. In this text $Y_k \sim \mathrm{Bin}(n, P(\boldsymbol{X}_k))$, and since the exponential family includes the binomial distribution the model considered is a generalised linear ditto.

With the notation used in eq. (5.3), the maximum likelihood (sometimes abbreviated ML) procedure is used for estimating the parameters $\boldsymbol{\theta} = [\beta_0, \dots, \beta_p]^T$, $p \in \mathbb{N}$. Let $\boldsymbol{y} = [Y_1, \dots, Y_n]^T$, $n \in \mathbb{N}$ denote the random vector of the response variable and denote the joint probability density function by $f(\boldsymbol{y}; \boldsymbol{\theta})$. Algebraically, the likelihood function $L(\boldsymbol{\theta}; \boldsymbol{y})$ is equivalent to $f(\boldsymbol{y}; \boldsymbol{\theta})$, but the emphasis is switched from $\boldsymbol{y}$, with $\boldsymbol{\theta}$ fixed, to the converse. The value $\widehat{\boldsymbol{\theta}}$ that maximises L is called the *maximum likelihood estimator* of $\boldsymbol{\theta}$. Maximising the likelihood function is equivalent to the often computationally less demanding task of maximising its logarithm, the *log-likelihood function* $l(\boldsymbol{\theta}; \boldsymbol{y}) := \ln L(\boldsymbol{\theta}; \boldsymbol{y})$. Thus, the maximum likelihood estimator is

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Omega_\Theta}{\arg\max}\, l(\boldsymbol{\theta}; \boldsymbol{y}), \tag{5.5}$$

with Ω_Θ as the parameter space. It is to be noted that maximum likelihood estimators have the *invariance property*, i.e. the maximum likelihood estimator for any function $g(\boldsymbol{\theta})$ is $g(\widehat{\boldsymbol{\theta}})$.

A matter to be considered when performing logistic regression, is if to use the conditional (L_C) or the unconditional (L_U) algorithm for calculating the likelihood function. The ratio between the number of explanatory variables, p, in the model and the number of observations, n, in the study is what determines which approach to apply. As a general rule, the unconditional formula is advantageous when the number of explanatory variables is small in comparison to the number of observations, i.e. $p \ll n$, and vice versa for the conditional formula. There is no exact definition of what is small and what is large in this context; the analyst will simply have to choose if the ratio does not belong to either extreme. In the study in this report, however, it is immediately obvious that the unconditional approach is preferable since the number of applicants (observations) is huge. The unconditional formula describes the joint probability of the study data as

$$L_C = \prod_{k=1}^{n_0} P(\boldsymbol{X}_k) \prod_{k=n_0+1}^{n} [1 - P(\boldsymbol{X}_k)], \quad n_0, n \in \mathbb{N}, \tag{5.6}$$

which in words is the product of the joint probability for the cases $k = 1, \dots, n_0$ where the response variable is true, $R = 1$, and the joint probability for the cases $k = n_0 + 1, \dots, n$ where the response variable is false, $R = 0$.

After the parameters of the model have been estimated, the fit and adequacy of the model remains to be determined. One way of achieving this is to compare it with a *saturated model*, i.e. a model with the same distribution and link function as the derived model, but where the number of observations, n, and parameters, p, are equal. If, however, r of the observations are replicates of each other, then the maximum number of parameters estimated in the saturated model is $m = n - r$. Let $\boldsymbol{\theta}_{max}$ denote the parameter vector for the saturated model, and with maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{max}$. An intrinsic property of L is that the more parameters a model has, the better the fit to the data will be, i.e. $L(\widehat{\boldsymbol{\theta}}_{max}; \boldsymbol{y}) \geq L(\widehat{\boldsymbol{\theta}}; \boldsymbol{y})$, with $\boldsymbol{\theta}$ as the parameters in the model of interest from now on called the *null model*, which is similar to the R^2-property in multiple linear regression. The likelihood ratio

$$\lambda = \frac{L(\widehat{\boldsymbol{\theta}}_{max}; \boldsymbol{y})}{L(\widehat{\boldsymbol{\theta}}; \boldsymbol{y})} \tag{5.7}$$

serves as a way to assess the goodness of fit for the model. For the same reasoning as the one leading to eq. (5.5), in practice $\ln \lambda$ is used. Moreover, the fact that $2 \ln \lambda$ approximately has a chi-squared distribution with $k = m - p$ degreees of freedom, leads to the definition of the *deviance* or *log-likelihood statistic*

$$D = 2[l(\widehat{\boldsymbol{\theta}}_{max}; \boldsymbol{y}) - l(\widehat{\boldsymbol{\theta}}; \boldsymbol{y})], \tag{5.8}$$

making it evident that a smaller deviance indicates a better fit.

Since the deviance shares properties similar to that of the χ^2 statistic, a common test to assess goodness of fit is to compare the deviance with the χ^2_{m-p} value, where n is the number of observations in the sample and p the number of parameters in the model (including the intercept β_0).

5.3 Variable Transformation

The explanatory variables in the customer data set are of different types; they can be nominal, ordinal or continuous. As an example, the variable age is continuous, whereas others are simply dichotomous. A list comprising all the explanatory variables considered will not be published due to their sensitive nature. In those cases where the explanatory variable is not obviously categorical it is transformed into a *cumulative dummy variable*.

The transformation into cumulative dummy variables is done as follows. The original variable is split into several intervals and the score range is investigated for each. When two adjacent intervals have significantly overlapping score ranges the two intervals are merged. When no more intervals are merged one interval in either of the far ends of the spectrum is discarded and the rest is considered to be one dummy variable. Then the interval next to the previously discarded one is disregarded as well, and together the remaining intervals make up the next dummy variable. The procedure is continued until only the last interval remains. An example is shown in figure 5.1 for the explanatory variable age. Say that the first interval consists of all aged 18 to 22, the next interval all aged 23 to 27, etc. The first dummy variable may then comprise all aged 23 and more, the second all aged 28 and more. This way, all explanatory variables in the model will be categorical. These variables are used to train the model.

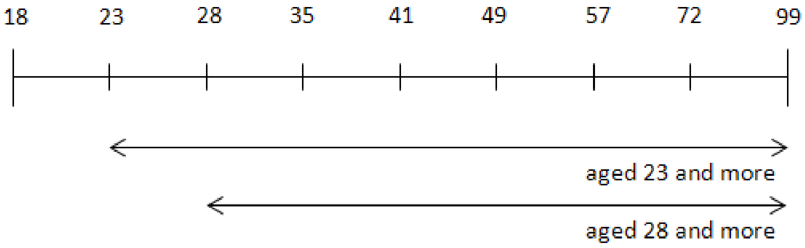

Figure 5.1: Example of cumulative dummy variables for the explanatory variable age.

5.4 Weight of Evidence and Information Value

Weight of evidence is often used in credit scoring as a measure of how well an attribute separates good from bad transactions, and in [1, p. 192] and [10, p. 81] it is defined as

$$W_i = \ln\left(\frac{\text{distr. of good}_i}{\text{distr. of bad}_i}\right) = \ln\left(\frac{N_i}{\sum_{i=1}^n N_i}\bigg/\frac{P_i}{\sum_{i=1}^n P_i}\right), \quad i = 1, \ldots, n, \tag{5.9}$$

where P denotes an occurrence (i.e. positive), N a non-occurrence (i.e. negative), i signifies the attribute under evaluation (e.g. an attribute of the explanatory variable age could be age less than 25 years), and n is the total number of attributes. A weight of evidence below zero implies that this particular attribute isolates a greater proportion of bad observations compared to good observations, and vice versa for a positive weight of evidence. Weight of evidence only accounts for the relative risk, it does nothing to address the relative contribution of each attribute. To this end, another measure called *information value* is used. Once again following [1, p. 193] and [10, p. 81] its definition is

$$\begin{aligned} I_V &= \sum_{i=1}^{n}(\text{distr. of good}_i - \text{distr. of bad}_i) \cdot \ln\left(\frac{\text{distr. of good}_i}{\text{distr. of bad}_i}\right) = \\ &= \sum_{i=1}^{n}\left(\frac{N_i}{\sum_{i=1}^n N_i} - \frac{P_i}{\sum_{i=1}^n P_i}\right) \cdot W_i, \quad i = 1, \ldots, n, \end{aligned} \tag{5.10}$$

where P denotes an occurrence (i.e. positive), N a non-occurrence (i.e. negative), i signifies the attribute under evaluation, n is the total number of attributes, and W_i is the weight of evidence. An intrinsic property of I_V is that it is always non-negative. The higher the value of I_V the better the predictive power of this explanatory variable.

5.5 Performance Measures

5.5.1 Gini Coefficient and AUC

When evaluating a case it is deemed to be either good or bad, and so either rejected or approved. Later on it is possible to determine which of these cases that were correctly classified: *true positives* (cases thought to be bad that were bad), *true negatives* (cases thought to be good that were good), *false positives* (cases thought to be bad, but they were good), and *false negatives* (cases thought to be good, but they were bad). Common practice is to refer to false positives as a *type I error* and to false negatives as a *type II error*. A graphical summary of the above is shown in table 5.2. In credit risk context the *false positive rate* is the rate of occurrence of positive test results in applications known to be good. This definition makes the false positive rate equal to 1 − *specificity* of the test. Similarly, the *true positive rate* is the rate of occurrence of positive test results in applications known to be bad, which explains why the true positive rate is also called the *sensitivity*.

	H_0 false (is good)	**H_0 true (is bad)**
Fail to reject H_0 (thought to be bad)	False positive	True positive
Reject H_0 (thought to be good)	True negative	False negative

Table 5.2: A table depicting the possible outcomes of hypothesis testing. The null hypothesis may be anything, but in credit risk context in this study the null hypothesis is that the application is bad.

Some way of assessing how well the scorecard is able to separate good from bad applications is required. For this the Gini coefficient will be used, and its definition is the area between the *receiver operating characteristic (ROC)* curve and the diagonal, as a percentage of the area above the diagonal. The x-axis of the ROC curve is the false positive rate and its y-axis is the true positive rate. Thus the ROC is represented by plotting the fraction of true positives out of the positives versus the fraction of false positives out of the negatives, and the Gini coefficient ranges from 0 (no separation between good and bad at all) and 1 (perfect separation). An example is shown in figure 5.2.

A measure akin to the Gini coefficient is the *AUROC* (Area Under Receiver Operating Characteristic) or, more commonly, the *AUC*, which is defined as the area under the ROC curve—exactly as its name implies. Equal to the AUC is the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. A model no better than a random guess would have an AUC of 0.5 whereas a value of 1 indicates the, unlikely, occurrence of perfect predictions. Similar interpretation applies to values less than 0.5 implying that the model is getting it wrong with some consistency, with 0 meaning perfectly wrong predictions. The AUC is related to the Gini coefficient, G_c, via the simple formula

$G_c = 2\text{AUC} - 1$. The above definitions are described more thoroughly in, e.g., [1, pp. 203-207].

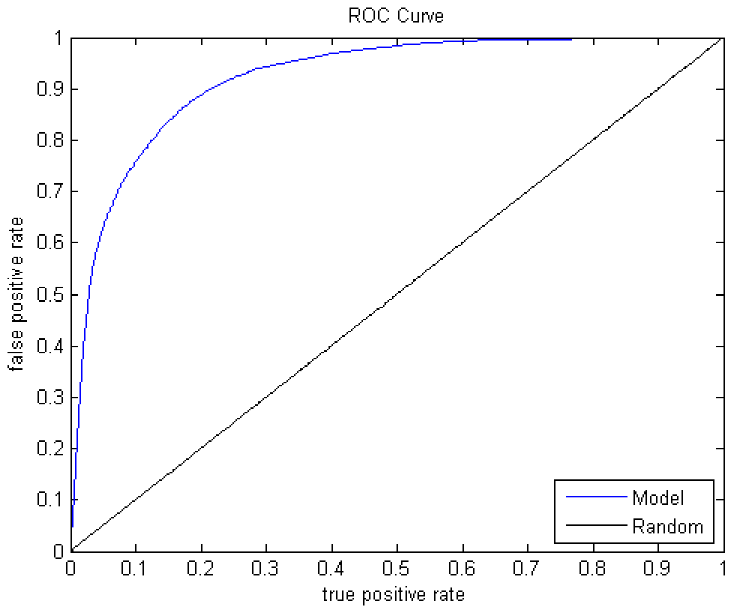

Figure 5.2: An example of the ROC curve from which both Gini coefficient and AUC is calculated.

5.5.2 Profitability

A different way to measure model efficiency is in terms of profitability—a fundamental property of most businesses. The main point here is that even if an order is paid in the end it must not necessarily generate a profit for the company, e.g. dunning costs could be large, and thus the purchase maybe should not be regarded as a good purchase. Because of the consumer being able to choose from a wide range of different payment possibilities the exact gain or loss per transaction is quite intricate to calculate. Hence, an estimation of the profitability is used in this study.

The models are compared on monetary losses where each transaction labelled as bad and getting a score over a certain *cut-off* value is modelled as adding its price to the total monetary loss of the model. At present the acceptance rate is around 85 percent and hence the cut-off value is chosen so that the 85 percent of the applications with the highest score are accepted in each set. The just described procedure is a simplification in several ways.

1. The loss of each transaction will not be equal to how much money the consumer is supposed to pay, since the company, e.g., has to add costs for send-

ing invoices and reminders or, possibly, subtract costs because the invoice may be partly paid.

2. A hard cut-off based solely on score is not the way it works at the company. When buying the customer has the option to choose from many different payment options, and depending on selected payment method a customer with a score normally rejected may be accepted anyhow.

3. Even though the purchase is labelled as bad after 90 days, as mentioned earlier, does not exclude the possibility that the customer will pay later. The probability of a *loss given default (LGD)* of the customer behind a purchase with the bad label decreases with a higher score. See, e.g., [9] for a more detailed coverage of LGD.

Despite the just listed shortcomings the estimated loss will give an indication of how profitable the model is.

The exact amount of the losses will not be presented because of that information being regarded as sensitive, instead the loss ratio

$$R_L = \frac{L_{ri}}{L_{wri}} \tag{5.11}$$

defined as the ratio between the estimated pecuniary loss with reject inference, L_{ri}, and the estimated pecuniary loss without reject inference, L_{wri}, will serve as a performance measure. If $R_L < 1$ a model with reject inference would yield a lower loss than one without, whereas if $R_L > 1$ a model with reject inference would yield a higher loss.

5.6 Model Selection and Validation

The final model is derived with *stepwise regression* based on the information value of the explanatory variables. The variable with the highest information value is added first to the model and checked if it provides a significant contribution, followed by the variable with the second highest information value, etc. This process is continued until the marginal information value of the variable to be added is negligible. In this report the limit of a marginal information value less than 10^{-6} is chosen since by then the possible improvements of the model are diminutive. The level of significance is set to 0.05. In each step it is also checked if any of the previously added variables no longer contributes significantly to the model; if so, that variable is removed.

A potential problem that can arise when training the model is that the end model may fit the training data set very well, but this does not necessarily mean that the fit would be as good for another data set that the model has not been trained on. To overcome this issue the model can be evaluated with the method of *cross validation* explained in [13].

The basic idea is what is implemented in the *holdout method*: to not use the whole data set for model training purposes, instead the data set is split into two parts, *training sample* and *holdout sample*, where the former is used to train the model and the latter is used to validate the derived model. The main problem with this method is that the outcome to a great extent may depend on how the random split between the two samples is made, i.e. the variance of the outcome may be quite high.

An improved method is *k-fold cross validation*, which in effect means that the data set is split into k subsets and where one of the subsets acts as the holdout sample and the remaining $k - 1$ subsets together form the training sample. The whole process is repeated k times. With this method the variance is decreased, and continues to decrease the higher the value of k.

In the model selection procedure a ten-fold cross validation procedure is implemented in each step of the process. When the best model has been chosen another validation is performed on the out of time validation set mentioned in section 4.1. A manual control of the included explanatory variables is made to ascertain that a variable that should add a negative contribution to the model really does so. This is a sanity check of the just performed number crunching. Furthermore, if one variable is missing that is thought to enhance the model this variable is added manually to check if it enhances the model. If the variable improves the model it is added.

Once the final model has been selected the deviance statistic described in section 5.2 is applied to assess whether there is evidence for a lack of fit at a level of significance of 0.05. If there is a lack of fit the model has to be revised.

6 Results

6.1 Histograms

In figure 6.1 histograms of the training set data of all, only good, and only bad credit applications, respectively, are shown for the model that incorporates reject inference. Figure 6.2 and figure 6.3 show the same type of histograms, but for the data from the validation and calibration set, respectively. Additionally, for the training set histograms of all accepted and of all rejected orders, respectively, are depicted in figure 6.4.

What is emphasised by figures 6.1, 6.2 and 6.3 is that the model manages to give bad applications another score distribution than good applications, even though the assigned score by no means is perfect; a distinct overlap of scores is seen. The calibration set proves to be the most difficult set in which to accurately distinguish between good and bad transactions, but this is no surprise since that set is chosen because it is as close as possible to a set of orders that were approved even though the initial scorecard would have rejected them. Recall the definition of the calibration set in section 4.1.

Figure 6.4 is of a slightly different nature; it gives a graphical representation of how the model derived with reject inference manages to distinguish accepted from

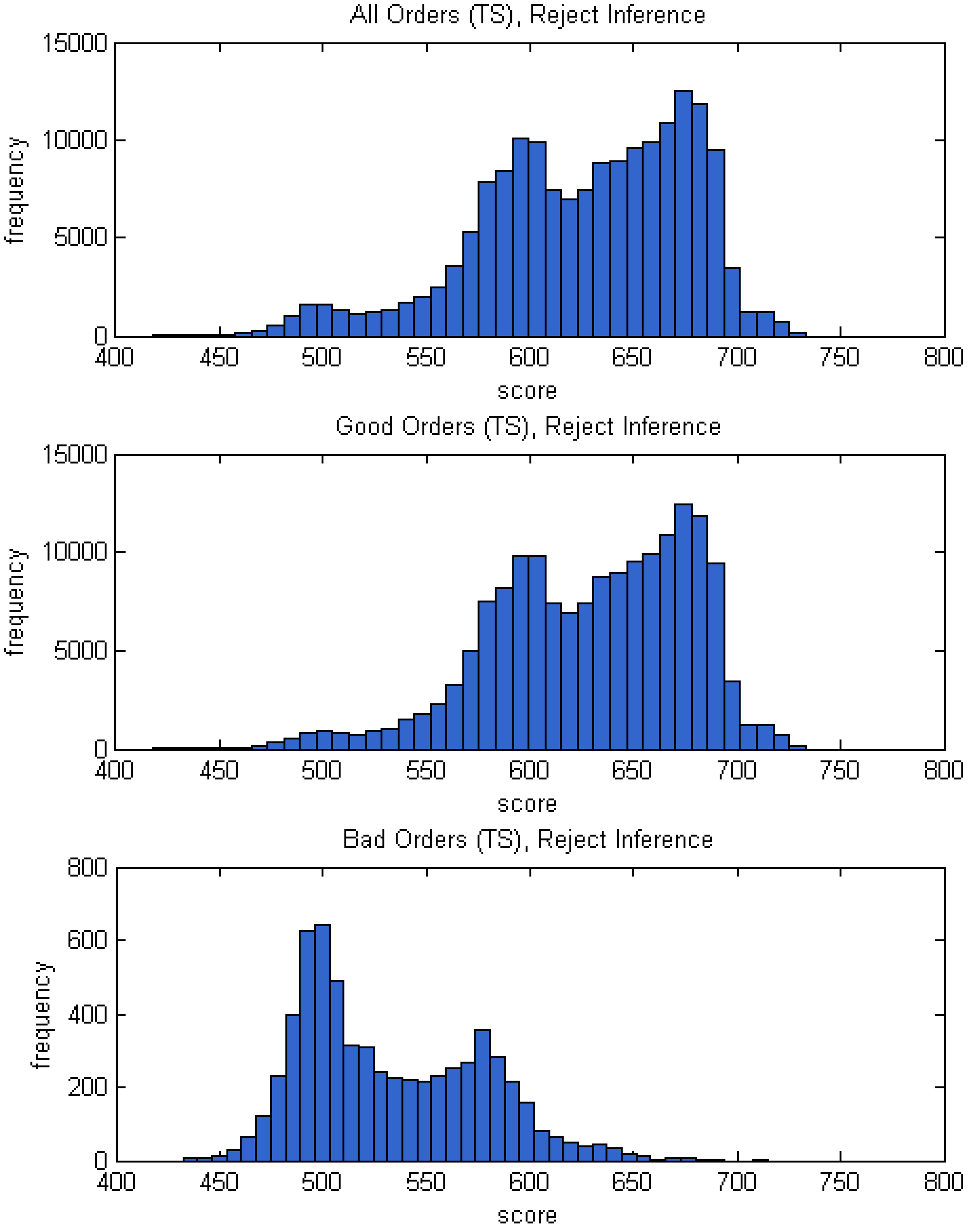

Figure 6.1: Histograms of the training set (TS) data scored by the reject inference model.

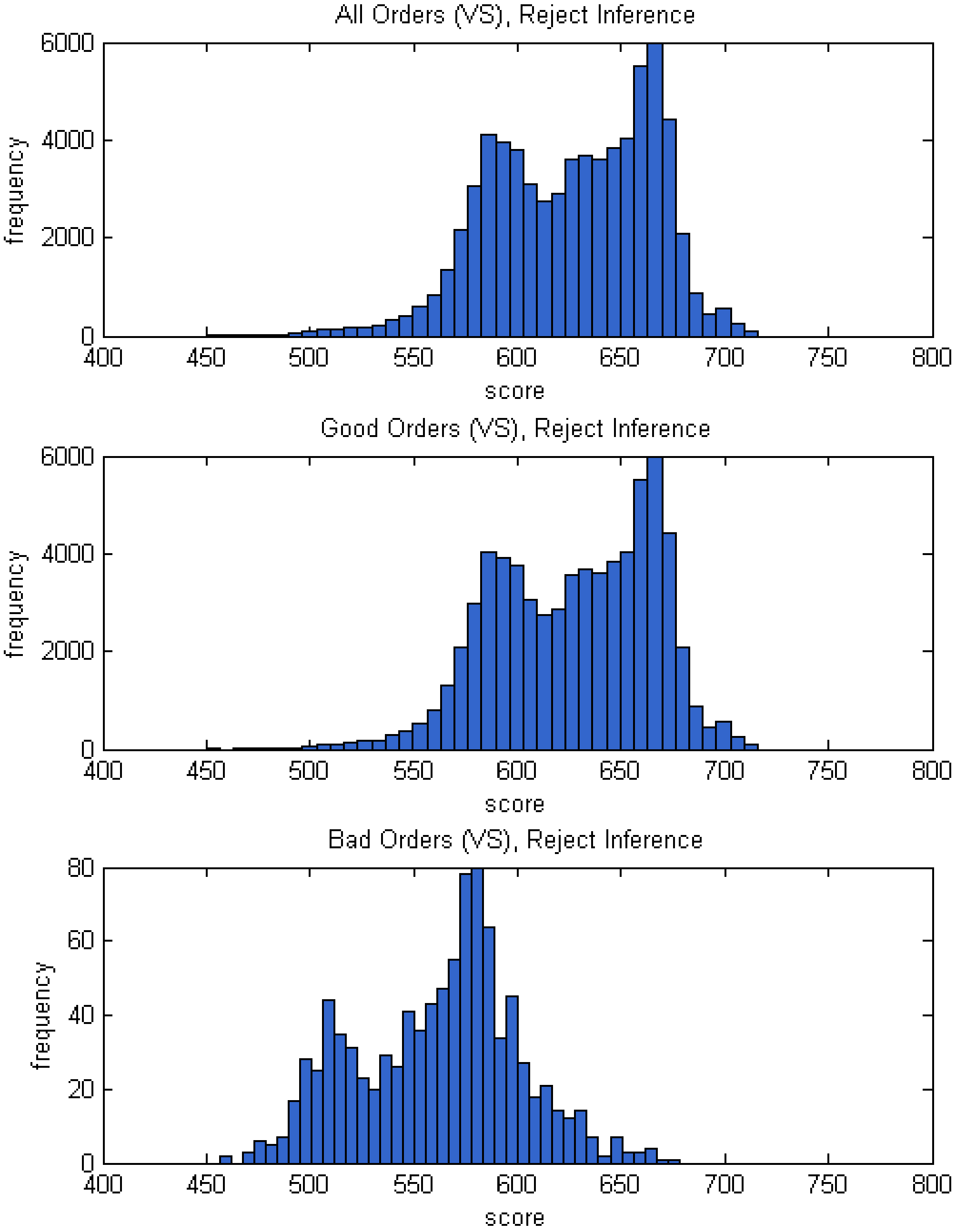

Figure 6.2: Histograms of the validation set (VS) data scored by the reject inference model.

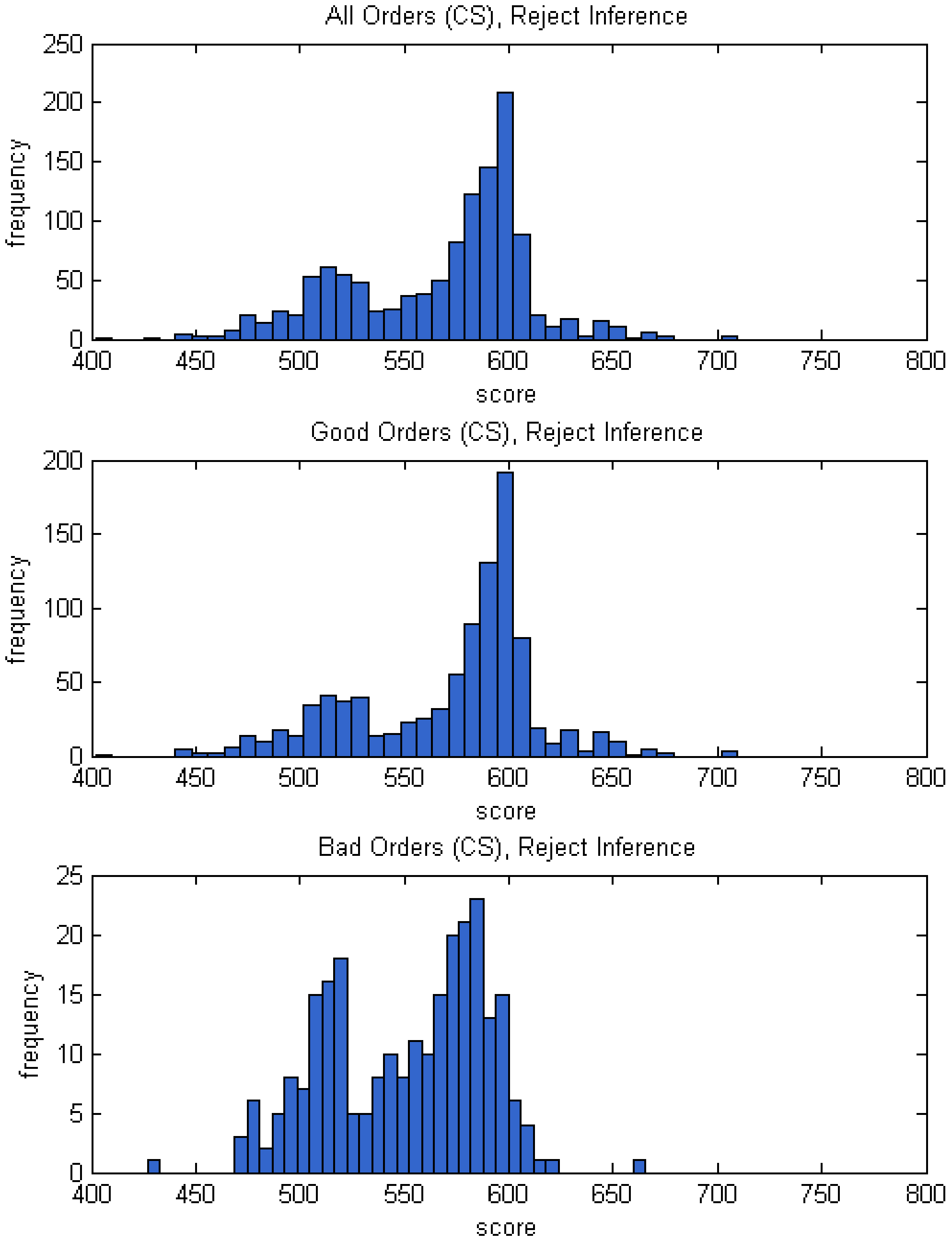

Figure 6.3: Histograms of the calibration set (CS) data scored by the reject inference model.

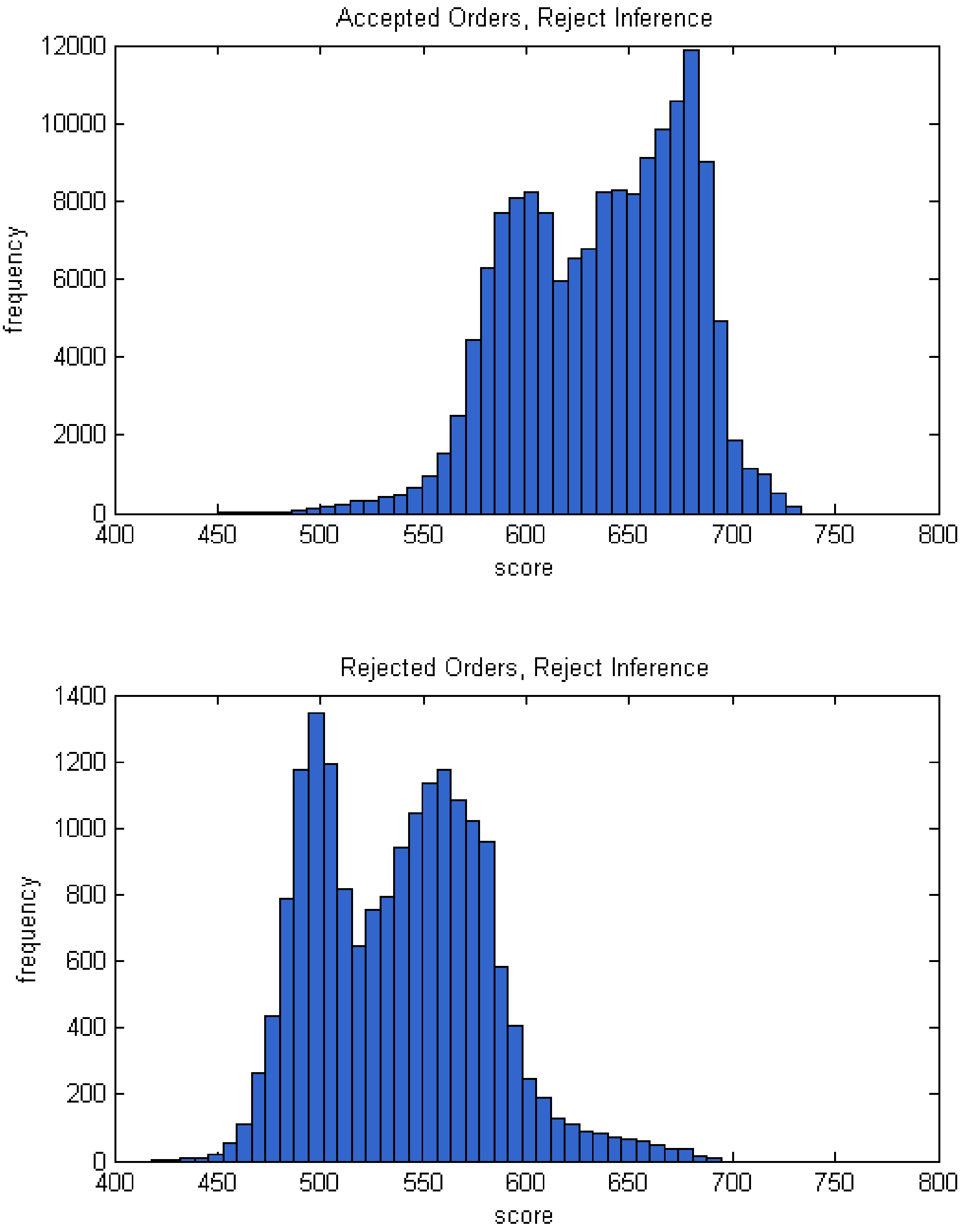

Figure 6.4: Histograms of the training set data scored by the reject inference model. The upper plot shows the accepted applications and the lower plot the rejected applications.

rejected orders. The outcome is that the newly derived model seems to agree with the old model: rejected orders get a lower score. It is important to note that the rejected/accepted label still is the one assigned by the old model.

As a form of check that the assumption of the calibration set being an approximation of rejects with a known outcome is not completely false, it is evident that the shape of the distribution of the rejected orders in figure 6.4 more closely resembles that of the calibration set in figure 6.3 than any of the other sets. However, it is far from a very good fit. The problem the models have with assessing the correct outcome of the orders in the calibration set is quantified via the Gini coefficient in the next section.

6.2 Gini Coefficient, ROC Curves and Profitability

The derived models with and without reject inference are compared on Gini coefficient for the training set, validation set, and calibration set in table 6.1 and on profitability, more specifically the estimated loss ratio, in table 6.2. As stated in section 3, because of the random assignment of reject outcome in the reject inference algorithm the altered input data will cause the results to fluctuate from run to run. The values shown for the reject inference model are therefore the mean of ten runs.

Model	$\text{Gini}_\mu\ (TS)$	$\text{Gini}_\mu\ (VS)$	$\text{Gini}_s\ (VS)$	$\text{Gini}_\mu\ (CS)$	$\text{Gini}_s\ (CS)$
RI	0.8429	0.7612	0.0035	0.3081	0.0197
No RI	0.7890	0.7582	0	0.2667	0

Table 6.1: Gini coefficients for the models, with and without reject inference, for training set (TS), validation set (VS) and calibration set (CS). Note that for the reject inference model the Gini coefficient varies from run to run, which is why the mean values are shown indexed by a μ. Furthermore, the standard deviation indicated by s is added to the table. The Gini coefficient values of the training set are of less interest because of overfitting issues.

	R_L	
	Mean	Standard Deviation
VS	0.9462	0.0680
CS	0.9060	0.1783

Table 6.2: The profitability of the models with and without reject inference is measured in terms of estimated loss ratio, R_L, for the validation set (VS) and the calibration set (CS). The values shown are the mean and standard deviation of ten runs.

Table 6.1 compares the models on their Gini coefficient telling how well the models separate bad from good applications. The Gini coefficient values of the training set are not as interesting as the other values since the former may be overfit to some degree, but the rather large difference between the two still gives an inkling that the predictive results of the model with reject inference may be slightly better. The Gini coefficients of the validation set varied very little between runs, which is shown by the small variance, but the difference in the Gini coefficients of the calibration set is a bit larger, also indicated by a higher variance. Regarding the mean values it is clear that there is only a very minor difference between the values of the Gini coefficient for the validation set, suggesting that correct assessment of accepted purchases is not enhanced particularly much by reject inference, but there is a slight improvement. The difference in Gini coefficient for the calibration set, i.e. orders with characteristics similar to rejects, on the other hand indicates that bad and good applications can be distinguished from each other to a higher degree with reject inference than without.

These results are corroborated by table 6.2 where the values of the estimated loss ratio, R_L, of the two models show that the losses in the validation set are lower with reject inference than without by a factor 0.946, and lower by a factor 0.906 in the calibration set. Once again the standard deviation is higher for the calibration set showing a greater spread between the values, but at the same time the mean is better for the calibration set than for the validation set. It is to be noted that there were observations of R_L higher than 1 in both sets. Thus, sometimes the random assignment of the parcelling algorithm worsens the outcome compared to a normal model without reject inference.

The just discussed Gini coefficients are calculated as the area under the ROC curve, see figures 6.5, 6.6 and 6.7. These figures elucidate the numbers in table 6.1 and show that for both the training and the calibration set the Gini coefficient is higher with reject inference, whereas for the human eye there is no discernible difference for the validation set. Because many different runs where made, and the figures look very much alike, only one reject inference model is depicted. The shown reject inference model has a Gini coefficient close to the mean of all.

7 Analysis

7.1 Analysis of Results

From the results in tables 6.1 and 6.2 it is seen that, on average, both the predictive power and the profitability of the model is enhanced when reject inference is applied. Regarding only the accepted purchases the improvement is not particularly large, but on the calibration set the difference is more easily discernible and it is to be noted that the improvement is aimed primarily at the type of orders that are rejected, i.e. those that the calibration set is supposed to represent.

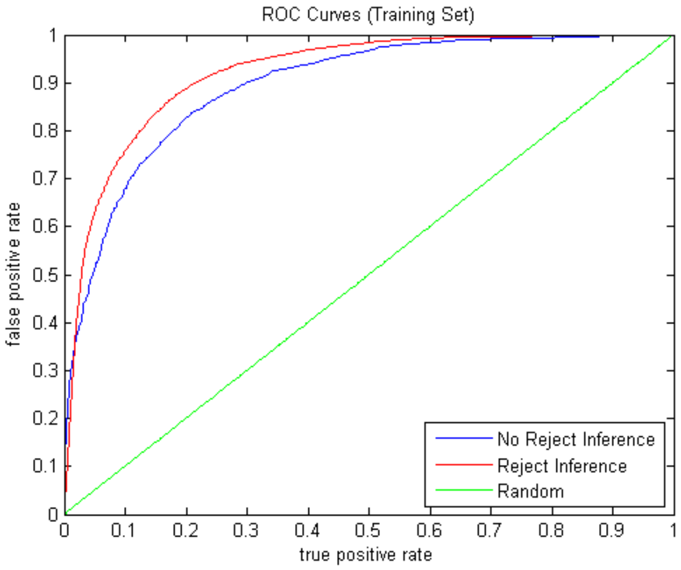

Figure 6.5: A plot of the ROC curves for the model with reject inference and for the one without reject inference. The straight line symbolises a model that assigns the good/bad label completely at random. Data is from the training set.

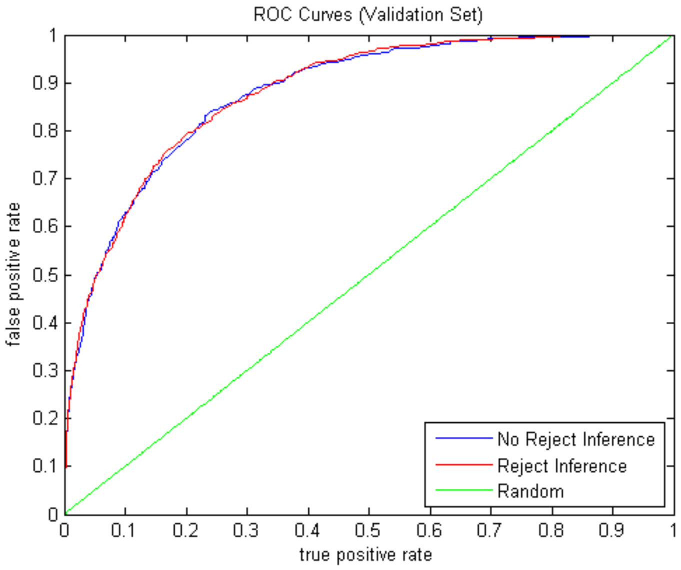

Figure 6.6: A plot of the ROC curves for the model with reject inference and for the one without reject inference. The straight line symbolises a model that assigns the good/bad label completely at random. Data is from the validation set.

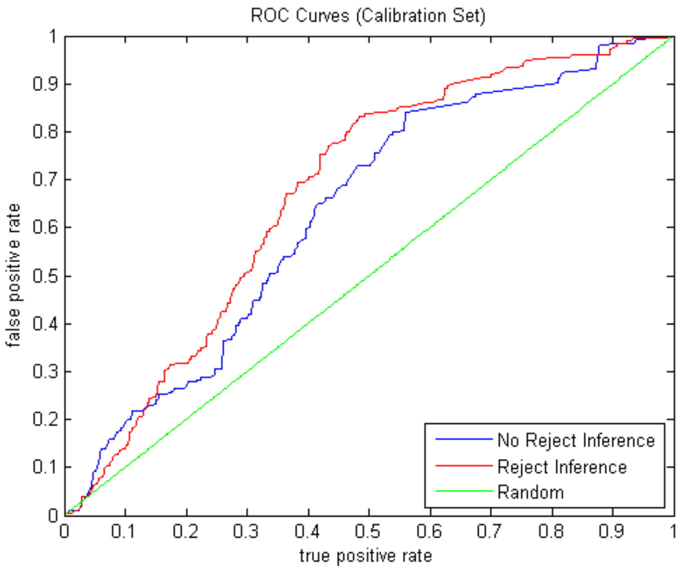

Figure 6.7: A plot of the ROC curves for the model with reject inference and for the one without reject inference. The straight line symbolises a model that assigns the good/bad label completely at random. Data is from the calibration set.

7.2 Analysis of Method and Assumptions

Previous analysis of reject inference by Crook and Banasik in [2, p. 24] concluded, inter alia, that useful results of reject inference primarily depend on precise estimation of the bad ratio, whereas elaborate tweaking of the model is secondary. These findings indicate that a likely source of error could be the estimation of the bad ratio in eq. (3.1) and in the assumptions preceding it. The exact dismemberment of the score range into several smaller intervals is one factor influencing the derived bad ratio, and its implementation is here mostly based on uniformity of the score ranges with the only exceptions, the two intervals in either end, to ensure that no intervals end up with very few or zero applications.

Optimal would be to have an extensive sample of rejected orders with known outcome, but with the available data this is unattainable. Hence the calibration set is used as a replacement, but its observations are suboptimal in several ways. One drawback is that these orders only can stem from a small subset of the total number of online stores that make up the total sample of the training set. See table 7.1 for a summary. The number of stores in the two sets are clearly very much unalike, but all of the stores in the calibration set range from big to huge in number of customers, so they still make up a sizeable part of the total number of orders in the training set. Another issue is the small size of the calibration set; in this study it merely comprises 1,218 purchases, many times less than the number of purchases in the training or validation set. The small size of the calibration set makes the result more prone to fluctuations. Increasing the time window to sample the data from would in turn increase the reliability of the results.

	Number of Stores
Calibration Set	13
Training Set	1,101

Table 7.1: A summary of how many stores the training and calibration set consist of, respectively.

Additionally, it has to be considered that the definition of an order being labelled as bad if it has not been paid within 90 days after due date is not perfect since there always are some customers paying later. A similar problem arises because some of the indeterminate orders eventually will be marked as bad when the real reason behind the non-payment or contestation is discovered.

7.3 Possible Extensions

In [8, p. 7] it is argued that the inherent randomness of the parcelling algorithm is a reason for concern. A better reject inference algorithm would be *fuzzy augmentation* that has a deterministic assignment of rejects to either the good or the bad label. Thus, implementing this reject inference technique instead of parcelling could be a way to get more stable results.

The time window of the data set has to be considered too, since e.g. seasonal variations or campaigns may influence purchase patterns and thereby the results. Another possible extension is to perform a stratified sampling of data from a much larger data set with a time window of, e.g., one year, and use this data set as input for training the model. Similarly, the validation set could comprise purchases from more diverse dates. A problem with a larger time window from is that since the company is growing and acquiring new customers regularly, too old data may misrepresent the customers of today.

The investigation could also be extended by trying to include the indeterminate observations and model them appropriately, since it is not unreasonable to assume that there are common behaviours amongst the customers behind these purchases. A way to mitigate the impact of the indeterminate orders is to extend the number of days from before a purchase is labelled as bad from 90 to a higher value.

8 Conclusion

The findings in this report indicate that a scoring model would benefit from incorporating reject inference. However, in order to validate the results and make them more general two modifications would be beneficial. One is to implement a deterministic reject inference procedure to stabilise the results, and the other is additional testing on a training data sample from a larger time window and with a calibration set with an increased number of observations.

9 Bibliography

References

[1] Anderson, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press, Inc., New York.

[2] Crook, J and Banasik, J. (2002). *Does Reject Inference Really Improve the Performance of Application Scoring Models?*. Working Paper Series No. 02/3, Credit Research Centre, The School of Management, University of Edinburgh.

[3] Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. Second Edition. Chapman & Hall/CRC, London.

[4] Feelders, A. J. (2003). *An Overview of Model Based Reject Inference for Credit Scoring*. Banff, Canada, Banff Credit Risk Conference 2003.

[5] Hand, D. J. and Henley, E. W. (1997). *Statistical Classification Methods in Consumer Credit Scoring: a Review*. Journal of the Royal Statistical Society. Series A (Statistics in Society), Vol. 160, No. 3 (1997), pp. 523-541. Blackwell Publishing for the Royal Statistical Society.

[6] Kleinbaum, D. G. and Klein, M. (2010). *Logistic Regression – A Self-Learning Text*. Third Edition. Springer, London.

[7] Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

[8] Montrichard, D. (2008). *Reject Inference Methodologies in Credit Risk Modeling*. Paper ST-160. Sesug, Inc.

[9] Schuermann, T. (2004). *What Do We Know About Loss Given Default?*. Working Paper Series, Federal Reserve Bank of New York, New York.

[10] Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. John Wiley and Sons, Inc., Hoboken, New Jersey.

[11] Verstraeten, G. and Van den Poel, D. (2004). *The Impact of Sample Bias on Consumer Credit Scoring Performance and Profitability*. Working Paper 2004/232, Faculty of Economics and Business Administration, Ghent University, Belgium.

[12] *Prepared Statement of The Federal Trade Commission on Credit Scoring before the House Banking And Financial Services Committee, Subcommittee on Financial Institutions And Consumer Credit*, Washington, D.C., September 21, 2000.
Source: http://www.ftc.gov/os/2000/09/creditscoring.htm.
Retrieval Date: 2012-02-14.

[13] *Cross Validation* (1997). An instructional text from Carnegie Mellon School of Computer Science.
Source: http://www.cs.cmu.edu/~schneide/tut5/node42.html.
Retrieval Date: 2012-05-03.