Process Scoring
for
Micro Credit Loans

Jie-Men Mok

Thesis (non-confidential)

November 2008

Process Scoring for Micro Credit Loans

Jie-Men Mok

Thesis (non-confidential)

November 2008

VU University
Faculty of Exact Sciences
Department of Business Mathematics and Informatics
De Boelelaan 1081a
1081 HV Amsterdam
The Netherlands

Host organization:
ING Retail Netherlands
Customer Intelligence
Multi Channel Campaign Management Debet
Location ACT. B 02 133
Bijlmerdreef 24
1102 CT Amsterdam
The Netherlands

Preface

Business Mathematics and Informatics (BMI) is a multidisciplinary programme which is aimed at improving business processes by applying a combination of methods based upon mathematics, computer science and business management. To gain work experience, the BMI programme is completed with an internship at an external organization. This thesis will report the course and the findings of my internship project at ING.

ING supports microfinance institutions of which one of them is the Opportunity Bank of Malawi (OIBM) which is located in Africa. The goal is to develop a credit scoring model for OIBM. This project has been initiated by Angenita de Rijke, ING advisor. I would like to thank her for her guidance and advices throughout the project. Furthermore, I would like to thank my colleagues for their support and insight regarding the project. Also, my thanks go to my academic supervisor, Marianne Jonker for her guidance and support.

Jie-Men Mok

Amsterdam, November 2008

Management summary

The goal of the project is to increase the efficiency of the application process of the IMC and SME loan by means of credit scoring. But data analysis indicates that this will be difficult, because the rejected applications are not entered in the system, the accepted applications with a good loan performance are not retained in the system and OIBM is not full grown enough to gather enough information about bad loans.

The credit committee accepts all the applications recommended by the loan officer. But when the applicant and/or the feasibility of the application is dubious, then the credit committee can decrease the credit risk by lowering the loan amount and/or the repayment term. Therefore, the target of the model changes to predicting application alteration. When the model estimates a high probability that the recommended application will not be altered, then the application can be delegated from the credit committee to the supervisor.

[Remainder classified]

Contents

1 Introduction

In the year 2001 the World Bank announced that there are 2.8 billion people of the world's 6 billion living on less than US$2 a day and 1.2 billion on less than US$1 a day. Of the people living on less than US$1 a day, 44% are in South Asia, 24% in Sub-Saharan Africa and 23% in East Asia and the Pacific. The World Bank had also noticed that poverty was rising in South Asia and Sub-Saharan Africa. This is why the World Bank had chosen for the theme 'Attacking Poverty' in its development report. One of the means to reduce poverty is microfinance.

This chapter will first give a description of microfinance. Then the background information of the project will be given and the problem description will be indicated.

1.1 Microfinance

People with low income lack financial capital to start up or to invest in their business. Microfinance is a way to provide financial services to the economically disadvantaged who have access to investment opportunities.

1.1.1 History

Microfinance was brought under attention by Muhammad Yunus in the seventies. The concept of microfinance originated long before the seventies but Yunus is famous for successfully executing microfinance to alleviate poverty by means of the Grameen Bank. In 2006, the Bangladeshi banker and economist Yunus and the Grameen Bank were awarded the Nobel Peace Prize for their effort to create economic and social development.

In 1974, Professor Yunus led his students of the Chittagong University on a field trip to the extremely poor households in the village Jobra. They interviewed a woman who made bamboo furniture. She had to borrow money from loan sharks against extremely high interest rates to finance the purchase of bamboo. Yunus discovered then that providing very small investment loans against fair interest rates could make a significant difference to a poor person. He tried to find banks that were willing to provide micro loans to the poor but was rejected due to the false image that the poor are bad clients.

The first loan was therefore financed by him self and consisted of US$27, which was made to 42 women in Jobra who made a profit and also repaid the loan quickly. Later, a bank was willing to cooperate with Yunus in giving out micro loans on the condition that he guaranteed every loan. It became a huge success and in 1983 he officially formed the Grameen Bank, meaning village bank, which is based on trust and solidarity. Currently, the Grameen Bank has more than 2,500 branches providing services to over 97% of the total villages in Bangladesh. It has almost 7.5 million borrowers of whom 97% are women.

1.1.2 Microcredits

Traditional credit loans require valuable assets to guarantee repayment of the loan. But microfinance institutions are adapted to the poor in a way that granting credit is based on trust. The credibility of the loan applicant is estimated by looking at the three 'C's: character, capacity and collateral. The character of an individual is judged by assessing the willingness of the applicant to repay the loan. The applicant should not only be willing but also be able to repay its loan. Hence the income of the applicant is used to indicate the capacity of an applicant. The collateral of an applicant is less crucial but also important since it provides a psychological pressure to repay. Usually the collateral is not valuable enough to cover the loss in case of default.

The credits handled in microfinance are so small that they are named microcredits. The loan amounts are usually between the US$5 and US$500. These are very small amounts to the people in developed countries but to the extreme poor a small amount can make a huge difference. Some microfinance institutions provide other financial services besides microcredits, like savings facilities and insurance. Usually the poor save on a non-financial way like purchasing more inventories. This type of saving has the drawback that the inventory can perish or be stolen. Another example of financial service is insurance. An insurance product can provide the opportunity to cover for example against natural disasters which can cause life threatening situations to the poor.

A microfinance institution can only be sustainable if the default rate is minimal, it must not exceed the 5%. It is therefore of great importance that the loan officer emphasizes the necessity to repay the loan on time. But the loan officer can also offer the borrowers enough flexibility to extend repayment in case of unforeseen circumstances. The main costs of a microfinance institution are the loan officers which are necessary to reduce the risk of default. The loan officer must maintain a close contact with the borrower by frequent visits in order to compensate for the absence of collateral. These costs are included in the interest rates of the borrower and this explains the higher rates compared to the commercial banks. Usually the interest rates of local money lenders are even higher than the rate of microfinance institutions.

1.1.3 Goal

The main target group of microfinance institutions is women because statistical results have indicated that they are more trustworthy in repaying their loan. Women are also seen as the heart of the household. They will insure that the earnings will be used to support the family and to create better opportunities for the next generation. Women in developing countries are usually placed in a submissive position. But through microfinance the status of women in their household can be elevated. By gaining more responsibility they will also gain more respect from their partner.

Microfinance has the same goal as charity, but it can also be distinguished from charity. When granting donations without additional conditions, the poor will not always use the gifts to create financial opportunities. Eventually, they can become dependent on constant aid which will only help them in the short term. But by providing microcredits, they will be assisted in making good investments. Microfinance offers the less fortunate the possibility to create better social circumstances.

1.2 Background information on this project

The year 2005 was declared by the United Nations to be the international year of microfinance to increase awareness and understanding of microfinance all over the world. ING was one of the head sponsors that year. ING wants to aid in providing better access to financial products and services to small entrepreneurs and individuals in developing countries. An important instrument for this is microfinance. Therefore, the program ING Microfinance Support was established to spread more knowledge by giving lectures about microfinance.

ING Microfinance Support also dispatches ING-specialists as consultants to microfinance institutions all over the world. These technical assistants share their knowledge and expertise with institutions so that the institutions can develop and expand their services. One of the microfinance institutions that received technical assistance is the Opportunity International Bank of Malawi (OIBM).

1.2.1 Malawi

OIBM is located in Malawi which is a country in south eastern Africa. Malawi is one of sub-Saharan Africa's most densely populated countries. The capital Lilongwe has a population exceeding 400,000 but Blantyre is the largest city in Malawi with nearly 500,000 inhabitants. Malawi has a gross national income per capita of US$170 per year which makes Malawi one of the poorest countries in the world. This is due to lack of mineral resources and harbours.

The residents are heavily dependent on agriculture. The most important export products are tobacco, tea and sugar. Malawi's reliance on agricultural products makes it vulnerable to external influences such as a declining demand of cigarettes and drought. Malawi has a subtropical climate with rain seasons from November to April. The residents have a low life expectancy of 43 which is mainly caused by HIV infection. The currency unit of Malawi is Malawian kwacha (MK) and €1 is approximately MK 200.

1.2.2 OIBM

The OIBM is a member of Opportunity International which is a global microfinance network. OIBM opened its first branch in the capital Lilongwe on May 23, 2003. The goal of the bank was to provide access to financial services for the economically disadvantaged people of Malawi, beginning with savings and later micro lending. In October 2007, OIBM has 3 head branches, 4 satellite branches and a mobile bank which is a bus that visits the rural areas.

The vision of OIBM is to help those living in poverty transform their lives through innovative, customer-driven financial services as a strong commercial microfinance bank. OIBM wishes to operate in a commercial way to guarantee endurance to their clients. Their core values are respect, integrity and commitment to the poor. OIBM has also shown to be inventive in supporting the poor when official identification, like passport or driving licence, turned out to be far too expensive for clients to acquire. In response to that, OIBM is using fingerprint identification technology for identifying clients.

1.2.3 ING support

In the last few years, OIBM has grown rapidly and is also confronted with competitors. In order to compete with other banks and to accommodate the growth, it is necessary to operate in a more efficient way. As a result, OIBM made a request for technical assistance from ING which has been accepted by Angenita de Rijke, senior researcher of the Customer Intelligence department within ING Retail Netherlands

The ING advisor made a visit to OIBM which lasted for two weeks. During the visit, the current situation of OIBM has been assessed and the requirements have been determined. Due to high work load and lack of time, she has not been able to execute the project herself. Therefore, the project has been handed over to me, under guidance of Angenita. The following section will present the problem description of the current project.

1.3 Problem description

The current loan application process takes too much time. OIBM would like to make the application process more efficiently by introducing credit scoring. OIBM would also like to have a clear overview of the loan portfolio and get a better understanding of the client by creating profiles of good and bad clients.

The required datasets for credit scoring have already been gathered and retrieved from the banking system of OIBM but the structure and contents of the datasets are unclear. This is why the target and the scope of the project will be further defined after the datasets have been analysed.

1.4 Outline of this document

The loan products and application process of OIBM will be described in chapter 2 in order to gain more background knowledge of the datasets. The banking system of OIBM and the given datasets will be explored in chapter 3. When the available possibilities of the data are known, the model will be described in chapter 4. In chapter 5, the given datasets will be processed into an analysis file on which the model will be built on.

The literature study of the mathematical techniques and methods which will be used when building the model, will be given chapter 6. Before starting to build the model, the analysis file will be explored in chapter 7 by means of the correlation matrix and the contingency table. Then, cluster analysis will indicate the presence of similar applications in chapter 8. Finally, the model will be built with the CHAID tree and logistic regression in respectively chapter 9 and 10. In the end, the conclusion and recommendation will be given in chapter 11.

The analysis will be performed by means of the statistical analysis software program SAS (version 8.2) and the corresponding SAS-code will be given in Appendix I: SAS code. Only the CHAID tree will be built in SPSS AnswerTree (version 3.0).

2 Preliminary exploration

First the product portfolio of OIBM will be described and then a selection of the loan products will be made on which the scoring model will focus. Afterwards, the application process of the selected loan products will be described. The preliminary exploration has already been executed by Angenita de Rijke, and therefore the description of the product portfolio and the application process is cited from her first report to OIBM *'First step in credit scoring'*.

2.1 Products

The main part of the portfolio of OIBM consists of several loan products and savings accounts. The savings account for the majority of the portfolio. The main loan products are divided in three categories:

- microcredit loans
- corporate and SME loans
- consumer loans

These three categories will be described in this section. A brief description of the other products can be found in section 2.1.4.

2.1.1 Microcredit loans

There are two types of microcredit loans: group loans (PTB: private trust banks) and individual micro credit (IMC). Both type of loans focus on the same sort of criteria for approval, which are called the 'three C's':

- Character: the willingness to repay the loan.
- Capacity or cash flow: the ability to repay.
- Collateral: the security provided by the borrower. For the lower loan amounts, this 'security' is mainly psychological, like household items and not registered properties. For higher amounts, houses and cars (comprehensively insured!) can be used. However, it is difficult to cash on the collateral when a client defaults.

Group loans are given to groups of at least seven members and maximal ten. Each group has a chair(wo)man and a treasurer. The treasurer collects all the amounts due and deposits them at the bank. The group has to cover for defaulting members; if the group defaults as a whole, then none of the members will be eligible for a loan any more. Group members can drop out when they default and the other members do not allow membership any more or when they move their business into another area.

Clients are recruited for IMCs in several ways:

1. Walking clients: they come into the office and ask for a loan.
2. Door-to-door marketing: loan officers visit business in their zones.
3. Sensthization meetings: meetings in the area, where people are informed about banking activities. These meetings are organized by cooperation with the local chief.
4. Referrals from clients.
5. Graduation from PTB-groups.

2.1.2 Corporate and SME loans

As the name says, the corporate and SME loans are given to larger enterprises. The lower SME loans are quite similar to the IMC loans, except for the limit. For these loans, no savings security is required. The businesses are much more 'grown up' and have substantial collateral, like:

- bill of sale: agreement in which significant assets of the client are listed (TV, computer, expensive household items).
- house
- vehicle.

The repayment period depends on the purpose of the loan: up to 12 months for assets and acquisition, possible longer for investments.

The financing of SME-3 customers (about 39) can hardly be called 'micro'. OIBM does not market them actively, but appreciates doing business with them when they approach OIBM. These firms have either grown from SME-2 or they have the same philosophy about doing business in Malawi and therefore select OIBM as their bank. According to Steve Mgwadira, the clients are loyal because of OIBMs flexibility and the initial opportunity OIBM provided for them while taking the first loan. Where competitors focus on collateral, OIBM assesses cash flow as the main criterion for acceptance.

2.1.3 Consumer loans

OIBM offers two types of consumer loans: mphamvu loans and employee loans. Mphamvu loans are meant for clients with a savings account who experience an emergency situation, like a funeral. Instead of using the savings account to pay for this emergency, the client is offered a mphamvu loan. This may seem strange, but discipline for saving is difficult in Malawi. Repaying a short term loan is easier, because it is an obligation to the bank. By using the loan for the emergency, the long term savings are protected. The assessment of the application for a mphamvu loan is based on the savings balance and the repayment history of previous loans. The security is provided by the savings account, which will be a fixed deposit during the loan.

Employee loans are a cooperation between OIBM and larger companies. Employees often ask their employer for a loan to see to their financial needs. To prevent spending too much time and especially capital, companies work with OIBM to supply these loans. The companies assess the client, determine the loan amount, deduct the monthly payments from the wages and settle the accounts when the employee leaves the company. OIBM just processes the applications and provide the capital. So for these loans, OIBM does not assess the client, but assesses the company.

2.1.4 Portfolio description

Besides the products described in section 2.1, some other products are available:

- Overdraft: only for large corporates or for very short time (2 months or less) facilities.
- Money market investments and deposits: money OIBM has invested / deposited with other financial institutions.
- Premium accounts: savings accounts with which the client should give notice in advance for a withdrawal (7 or 21 days).
- FDCA: fixed deposits in a foreign currency.

[Remainder classified]

2.2 Application process

This project will focus on the IMCs. Therefore the application process will be described from an IMC point of view. For SMEs, which will be the next step, additional comments will be made in a separate section.

Figure 2.1 provides an overview of the application process in general. In each step of the process, an accept/reject decision is made, where 'accept' means: proceed to the next step. The person making this decision, is given in the right column.

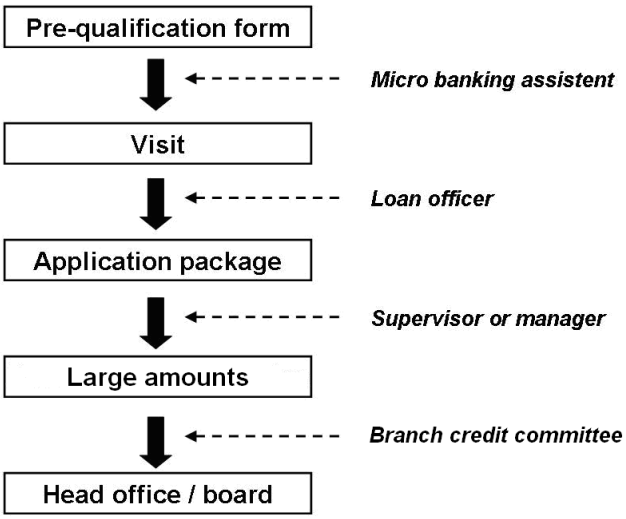

Figure 2.1 Overview of the application process in general

[Remainder classified]

3 Data exploration

[Classified]

4 Model description

[Classified]

5 Analysis file

[Classified]

6 Statistical methods

This chapter will describe the summary of the performed literature study on the relevant mathematical methods and techniques which will be used when building the models.

The correlation matrix will be described in section 6.1 which will be used to examine the dependencies between variables. The contingency table in section 6.2 will be used to investigate the direct relationship between the explanatory variables and the target variable. The cluster analysis in section 6.3 will indicate the presence of similar applications.

The application alteration model and the loan performance model will be built by means of the CHAID tree and regression analysis which are described respectively in section 6.4 and 6.5. The variable selection method in the regression analysis will be described in section 6.6. In section 6.7, the transformation of the explanatory variables will be described. The lift curve in section 6.9 and the ROC curve in section 6.10 can be used to evaluate the models. Section 6.12 shows how the cut-off point with optimal classification accuracy can be determined.

6.1 Correlation matrix

Pearson's correlation coefficient indicates the linear relationship between two random variables X and Y for the paired observations (X_i, Y_i), …, (X_N, Y_N) which are independent and identically distributed as X and Y. The correlation coefficient is defined as follows:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

$\rho_{X,Y}$ is estimated by $\hat{\rho}_{X,Y}$ as follows:

$$\hat{\rho}_{X,Y} = \frac{\sum_{i=1}^{N}\left(\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right)}{\sqrt{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2}}, \text{ with } \overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i .$$

The correlation coefficient lies within the range of -1 and 1, where 0 represents no linear relationship and 1 or -1 represents a strong relationship.

The correlation matrix denotes the correlation coefficient between multiple random variables in a matrix, see table 6.1 with random variables A, B and C. Note that the correlation of identical variables is 1 and that the lower left half of the matrix is mirrored in the diagonal line to the upper right half where $\rho_{A,B} = \rho_{B,A}$, $\rho_{A,C} = \rho_{C,A}$ and $\rho_{B,C} = \rho_{C,B}$.

	A	**B**	**C**
A	1	$\rho_{A,B}$	$\rho_{A,C}$
B	$\rho_{B,A}$	1	$\rho_{B,C}$
C	$\rho_{C,A}$	$\rho_{C,B}$	1

Table 6.1 Correlation matrix

The test of Pearson's correlation can be performed in order to examine the linear relationship between variables A and B. The null hypothesis H_0 and the alternative hypothesis H_1 and the statistic T are defined below:

$H_0: \rho_{A,B} = 0$
$H_1: \rho_{A,B} \neq 0$

$$T = \frac{\hat{\rho}_{A,B}\sqrt{N-2}}{\sqrt{1-\hat{\rho}_{A,B}^{\,2}}} \sim t\text{-distribution with } n-2 \text{ degrees of freedom under } H_0$$

At significance level α, if the p-value $\leq \alpha$, then H_0 is rejected. This means that the variables A and B have a linear relationship.

6.2 Contingency table

Contingency tables can be used to analyse the relationships between two categorical variables. An example of a contingency table for variables X and Y is shown in table 6.2, where the values of X is split in separate groups of X_1 and X_2 and Y in Y_1 and Y_2. The numbers in the final column and row show the marginal totals of the observations and the number in the right corner below shows the grand total.

		Y		
		y_1	y_1	Total
X	x_1	40	30	70
	x_2	20	10	30
	Total	60	40	100

Table 6.2 Contingency table

The contingency table can be used to perform the Pearson's chi-squared test which will be further described in section 6.4.1.

6.3 Cluster analysis

Cluster analysis groups the raw data into clusters. The clusters are formed in such a way that the differences within clusters are minimal and the differences between the clusters are maximal. These two objectives can be contradicting. The differences within clusters can be minimized by forming many clusters, but then the differences between the clusters are usually not maximal. The same holds for the other objective. The differences between the clusters are maximal when the number of clusters is low, but then the differences within the clusters are usually not minimal. Therefore, the right balance needs to be found by weighing the importance of the objectives.

The cluster analysis can be performed in many ways due to varying methods and techniques. There are two main types of clustering: hierarchical and partitional. The following sections will describe each cluster type.

6.3.1 Hierarchical clustering

In hierarchical clustering, the clusters are nested and can be organized as a tree, see figure 6.1. Each cluster node has cluster children, except for the leaf nodes. Each cluster node also has a cluster parent, except for the cluster root. When the number of clusters increases, then the cluster with the most dissimilar observations will be disjoint in separate clusters. But when the number of clusters decreases, then the most similar clusters will be merged together.

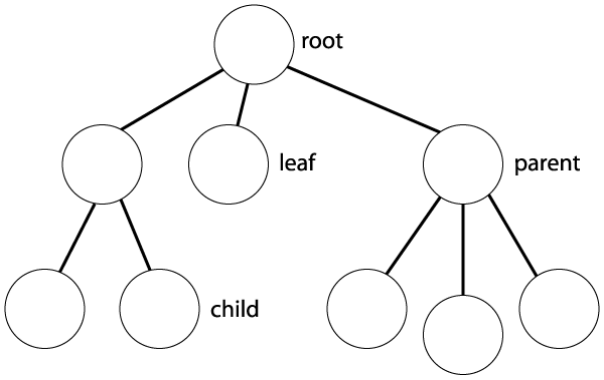

Figure 6.1 Tree

There are two approaches for hierarchical clustering: agglomerative and divisive. The agglomerative method starts at the bottom of the tree and classifies each observation as a cluster. In each step, the most similar clusters are combined together until there is only one cluster which contains all observations.

The divisive method moves in the reverse order of the agglomerative method. The divisive method starts at the top of the tree with one cluster containing all observations. In each step, every cluster is divided in separate clusters which are most dissimilar. The (dis)similarity between clusters can be measured in different ways and the clusters can also be joined or divided in different ways. These will not be further described since the hierarchical clustering technique has not been applied in this project.

The hierarchical clustering can be graphically reflected with a tree and is easy and intuitive. But this method does not globally optimize the objectives to maximize similarities within clusters and dissimilarities between clusters, because it is locally decided which clusters should be merged or split. Once the clusters have been merged or split, this can not be undone at a later time.

The major drawback is that the hierarchical clustering method is not suitable for large datasets. When the dataset contains too many observations, the algorithm will be slow and is not efficient enough to be handled by the current computers. Another weakness is that hierarchical clustering is sensitive for outliers and irrelevant variables.

6.3.2 Partitional clustering

In partitional clustering, the observations are grouped in disjoint clusters where each observation belongs to one cluster which is also called the *k*-means algorithm. See figure 6.2 for an example where the observations are two-dimensional.

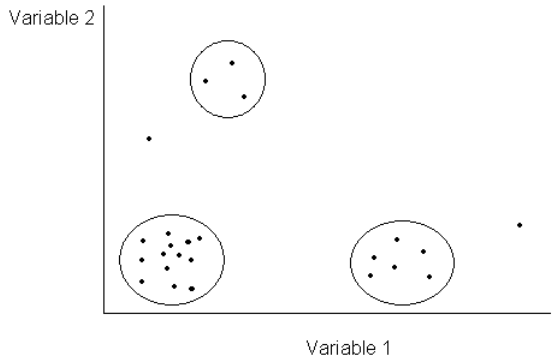

Figure 6.2 Partitional clustering

The *k*-means algorithm can be executed by following the steps below:

1. Define the number of clusters which is denoted by *K*.
2. Choose *K* initial seed points.
3. Assign each observation to the seed point where the similarity is maximal.
4. Recompute each seed point of the clusters.
5. If the seed points changes, then go to step 3. Otherwise, stop.

The similarity between the observations can be measured in different ways. The most frequently used is the Euclidean distance which should only be applied to quantitative variables. The Euclidean distance between observation x and y with p coordinates is defined as follows:

$$dist(x,y) = \sqrt{\sum_{k=1}^{p}(x_k - y_k)^2}$$

The Euclidean distance can be used to measure the similarity in step 3 and 4. In step 3, each observation is assigned to a seed point by minimizing the distance between the observation and the seed point. The seed point in step 4 can be determined with the least squares estimation. There are K clusters where cluster i contains m_i observations. The objective then is to minimize the sum of squared error (*SSE*) in the distance between each observation x_{ij} to the closest seed point s_i where $j = 1, \ldots, m_i$. The *SSE* is formulated below:

$$SSE = \sum_{i=1}^{K}\sum_{j=1}^{m_i} dist(s_i, x_{ij})^2$$

The seed point which minimizes the *SSE* of the cluster is the mean of the cluster, which is also called the centroid. This can be proved when the centroid is determined by setting the derivative of the *SSE* to zero, which is solved below for the one-dimensional centroid c_k for cluster k:

$$\begin{aligned}\frac{\partial}{\partial c_k} SSE &= \frac{\partial}{\partial c_k} \sum_{i=1}^{K} \sum_{j=1}^{m_i} \left(c_i - x_{ij}\right)^2 \\ &= \sum_{i=1}^{K} \sum_{j=1}^{m_i} \frac{\partial}{\partial c_k} \left(c_i - x_{ij}\right)^2 \\ &= \sum_{j=1}^{m_k} 2\left(c_k - x_{kj}\right) = 0 \Rightarrow m_k c_k = \sum_{j=1}^{m_k} x_{kj} \Rightarrow c_k = \frac{1}{m_k} \sum_{j=1}^{m_k} x_{kj} = \bar{x}_k, \text{ where } k = 1,..., K\end{aligned}$$

Instead of taking the mean of the cluster, it is also possible to take the median of the cluster which is called the medoid. The median is calculated by taking the middle of the observations which are ranked from low to high.

The minimalization of the *SSE* is executed in step 3 and 4 of the *k*-means algorithm. In step 3 the clusters are formed by assigning observations to the nearest centroid and in step 4 the centroid are recomputed by further minimizing the *SSE*.

When there are outliers present, then the cluster centroids may not be representative. The *SSE* will be higher when the cluster contains outliers, because outliers have a relatively large distance from other observations. The high *SSE* will bias the cluster centroid. It is therefore useful to discover and remove outliers first.

In partitional clustering the number of clusters needs to be pre-defined, which is not the case in hierarchical clustering. But the partitional clustering is more efficiently than the hierarchical clustering and it can therefore handle large datasets. The partitional clustering is also able to optimize the objectives of maximum similarities and dissimilarities in a global way by minimizing the *SSE*.

When performing cluster analysis, it should not be considered as ‘science’ but more as ‘art’. There is more than one way to perform cluster analysis which will depend on the data and the purpose of the analysis. There are no correct solutions, only suggestions which indicate what the solution might be.

6.3.3 Standardization

Usually, the data is first standardized before performing cluster analysis. This is because the variables with larger magnitude will dominate. Then the large value for variance will bias the distance which is used to measure the similarity of applications.

The standardized value Z of the unstandardized value X with mean μ and standard deviation σ is calculated as follows:

$$Z = \frac{X - \mu}{\sigma},$$

where $Z \sim \text{Norm}(0, 1)$.

6.3.4 Missing values

It is possible that variables of the observations contain missing values. In SAS, when the first q variables are non-missings then the distance between an observation x with missing values and the centroid c is computed as follows:

$$dist(c,x) = \sqrt{\frac{p}{q}\sum_{i=1}^{q}(x_i - c_i)^2} \ ,$$

where

p = total number of variables
q = number of variables with non-missing variables
x_i = value of the i^{th} coordinate for the observation
c_i = value of the i^{th} coordinate for the centroid

This means that the distance of the observations is being compensated for the missing values with rate $\sqrt{p/q}$.

6.4 CHAID tree

The acronym CHAID stands for chi-squared automatic interaction detector, which is a decision tree technique. It is used to examine the relationships between the response variable and the predictors. The CHAID model does not only indicate the significant variables but also their interactions. The CHAID model shows how selecting certain values for a combination of predictors can influence the response variable.

The CHAID tree will be explained by means of an example shown in figure 6.3, where the response variable is located at the root of the tree. In the second level of the tree, the population is split in three separate groups indicating the marital status. The child nodes of the tree denote the selection in the first line, the response rate in the second line and the number of response observations in the final line. In the third level, the married ones are split by gender and the divorced ones by absence or presence of pets. The leaf nodes show the selection of the population with the corresponding response rates. The overall response rate of 10% has risen to 50% for the married and male population and for the divorced population without pets.

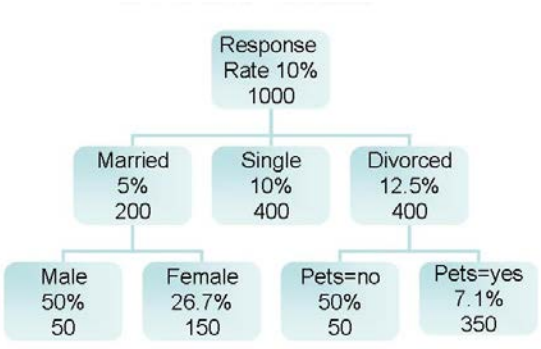

Figure 6.3 CHAID tree

The steps for the CHAID algorithm are as follows:

1. Select the response variable as the root of the tree.
2. If the predictor is continuous, create a categorical predictor by dividing the continuous predictor into a number of categories with an approximately equal number of cases.
3. Perform the chi-squared test of independence in order to indicate the difference between the allowable pair of predictor categories with respect to the response variable. The test of independence will be further described in section 6.4.1.
4. If there is no difference between the pair of categories, then merge the pair.
5. Select a new pair of (merged) predictor categories and go to step 3. If no new pair of categories can be selected, then the resulting set of categories is the best split with respect to the response variable.
6. Select the next predictor and go to step 2. If all predictors have passed through step 2, then select the predictor where the split of the predictor returns the best prediction of the response variable, and use the split to form the child nodes in the next level.
7. For each child node in the next level, go to step 2. Stop when one of the stopping rules is met:
    - None of the predictors can be split in different categories.
    - The maximum number of levels is reached.
    - The number of cases in the terminal node is less than the minimum number of cases for parent nodes.
    - When splitting the node in step 6, then the number of cases in one or more child nodes is less than the minimum number of cases for child nodes.

6.4.1 Chi-squared test of independence

The difference between the categories in step 3 is examined by performing the Pearson's chi-squared test of independence. The null hypothesis of the test is formulated as follows:

H_0: predictor is independent of response variable.

The chi-squared statistic X^2 for a dichotomous response variable is calculated as follows, where X^2 has a chi-squared distribution under H_0 with 2 degrees of freedom:

$$X^2 = \frac{(e_0 - o_0)^2}{e_0} + \frac{(e_1 - o_1)^2}{e_1},$$

where

e_i = # expected observations under H_0 where response category is i for $i = \{0, 1\}$
o_i = # number of observed observations where response category is i for $i = \{0, 1\}$.

When there has been assumed that the predictor is independent of the response variable, then each predictor category is expected to be distributed over the response categories in the same way as the overall distribution. This can be illustrated with a two by two contingency table, where the expected number E_{ij} for the cell in the i^{th} row and the j^{th} column is calculated as follows:

$$E_{ij} = \frac{R_i \times C_j}{T},$$

where

R_i = total number of observations in the i^{th} row for $i = \{1, 2\}$
C_j = total number of observations in the j^{th} column for $j = \{1, 2\}$
T = total number of observations in the whole table

The expected numbers for the example in table 6.2 are shown in brackets in table 6.3.

		Y		
		y_1	y_1	Total
X	x_1	40 (42)	30 (28)	70
	x_2	20 (18)	10 (12)	30
	Total	60	40	100

Table 6.3 Observed and expected numbers in contingency table

The chi-squared statistic X^2 is used to calculate the *p*-value. The *p*-value is $P_{H_0}\left(X^2 > x\right)$ for observed *x*. The null hypothesis is rejected if the *p*-value is smaller than or equal to the significance level.

6.4.2 Bonferroni adjustment

Each predictor variable in step 5 can be split in different ways. When multiple tests are performed, then the probability of making at least one Type I error increases. The Type I error is the error of rejecting the null hypothesis when the null hypothesis is in reality true. But when additional tests are performed, then it will become more likely that the null hypothesis will be rejected. This problem is called the inflation of the significance level.

The probability of making a Type I error for the combined tests is formulated below:

$$\alpha = 1 - \left(1 - \alpha_i\right)^n ,$$

where

α = overall significance level for all tests
α_i = significance level for individual test
n = number of tests.

The significance level is defined as the probability of making a Type I error. When the significance level is set at $\alpha_i = 0.05$, then the probability of making a Type I error for each test is also 0.05. Therefore, the probability of not making a Type I error is the complementary part, which is $(1 - \alpha_i) = 0.95$. When there are n events then the probability that all n events occur is the product of their probabilities. Therefore, if the number of tests is $n = 10$, then the probability of not making a Type I error for all tests is $(1 - \alpha_i)^n \approx 0.6$. This means that the probability of making a Type I error for all tests is $1 - (1 - \alpha_i)^n \approx 0.4$. In conclusion, when the α level is 0.05 for each of the 10 tests then the probability of incorrectly rejecting the null hypothesis is approximately 40%, which is far above the probability of 5% for the single test.

The probability of making a Type I error for the individual test has been rewritten below and is sometimes called the Šidàk equation:

$$\alpha_i = 1 - \left(1 - \alpha\right)^{1/n}$$

The Šidàk equation can be simplified by approximating it with the first order of the Taylor polynomial. Given that α is a probability, it is bounded as follows: $0 < \alpha < 1$. Therefore, the Šidàk equation can be rewritten as the binomial series formula, which is stated below:

$$(1+x)^k = 1 + kx + \frac{k(k-1)}{2}x^2 + \ldots, \quad \text{for } |x| < 1 \text{ with parameter } k$$

The last part of the right hand side of the Šidàk equation can be made equal to the binomial series formula with the following equations:

$$x = -\alpha$$
$$k = \frac{1}{n}$$

When replacing the preceding equations in the last part of the right hand side of the Šidàk equation, then the binomial series formula can be applied as follows:

$$\begin{aligned}(1-\alpha)^{1/n} &= 1 + \frac{1}{n}(-\alpha) + \frac{\left(1/n\right)\left(1/n - 1\right)}{2}(-\alpha)^2 + \ldots \\ &\approx 1 - \frac{\alpha}{n}\end{aligned}$$, where $0 < \alpha < 1$ and $n > 0$

Note that the terms in the right hand side equation become smaller and smaller. In the last step, the terms that come after the second term, can be dropped since they approach zero and are small with respect to the first two terms. When the approximation has been replaced in the Šidàk equation, this results in the Bonferroni adjustment:

$$\alpha_i = \frac{\alpha}{n}, \qquad i = 1, \ldots, n$$

The Bonferroni adjustment is less accurate than the Šidàk equation, but it is easier to compute. The Bonferroni adjustment can be applied on the significance level for the individual test of independence in step 3 of the CHAID algorithm.

When the significance level is corrected for multiple testing, then the significance level for the individual test decreases and becomes more strict. This will decrease the probability of making a Type I error but as a trade-off, the probability of making a Type II error increases. The Type II error is the error of not rejecting the null hypothesis when in reality the alternative hypothesis is true. Usually, the Type II error is less problematic than the Type I error.

The CHAID model can be evaluated by means of the lift curve and ROC curve, see respectively section 6.9 and 6.10.

6.5 Regression analysis

Regression analysis can be used to investigate the relationship between the target variable and one or more explanatory variables. When the parameters in the regression equation have been estimated, then this can also be used to predict the target variable.

The regression equation models the target variable Y as a function of the random component ε and the systematic component f, which is a function of the explanatory variables X and the parameters β:

$$Y_i = f(X_i, \beta) + \varepsilon_i, \quad \text{for observation } i = 1, \ldots, N$$

where

$X_i = (1, X_{i1}, \ldots, X_{iP})$
$\beta = (\beta_0, \ldots, \beta_P)$

The vector of parameters β are unknown constants where β_0 is called the intercept and β_1 to β_P are the regression coefficients. The error term is a random variable which represents the unobservable noise of the data. The objective of regression analysis is to estimate the parameters in such a way that it returns the optimal goodness of fit with the observations of the target and explanatory variables. It is expected that the estimated parameters returns the best prediction of the target variable given the observed values of the explanatory variables.

Regression analysis has the following underlying assumption:

- The future data samples must be similarly obtained as the training sample.
- The error is normal distributed with a mean of zero.
- In a linear regression model, the explanatory variables are linearly independent of each other, which means that it is not possible to express one variable as a linear combination of other variables. If a variable is a linear combination of the others, then this is called multicollinearity.
- The errors are independent.
- The variance of the error is constant throughout the observations, which is also called homoscedasticity.

It is usually recommended that the data sample should have at least ten to twenty times as many observations as the number of explanatory variables, otherwise the parameter estimates are probably unstable.

The relationship between the target and explanatory variables can be modelled in different ways which results in different types of regression. First, the linear regression will be described which will make it easier to explain the logistic regression.

6.5.1 Linear regression

In linear regression, the target variable is a linear combination of the regression coefficients but it does not have to be linear in the explanatory variables. When the target variable Y is linear in both the regression coefficients β and the explanatory variables X, then the linear regression equation is formulated as follows:

$$Y_i = \beta_0 + X_{i1}\beta_1 + \ldots + X_{iP}\beta_P + \varepsilon_i, \quad \text{for } i = 1, \ldots, N \text{ with } P \text{ variables}$$

where the error term ε has a zero mean:

$$E\varepsilon_i = 0$$

The expected target variable is estimated by the linear regression line, which has been illustrated in figure 6.4 where there is only one explanatory variable. The linear line represents the least squared estimator.

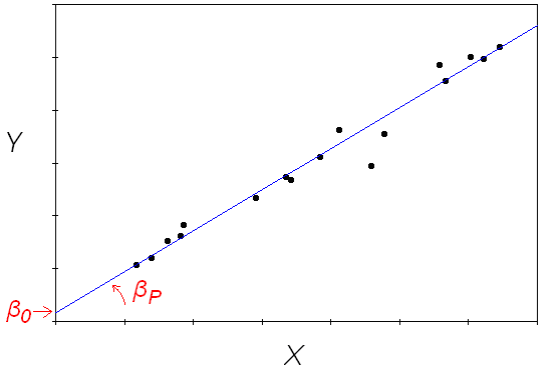

Figure 6.4 Linear regression line

The regression line estimates the data points where the line starts at the intercept and the regression coefficient represents the slope of the line. The regression line is defined as follows:

$$E[Y_i] = \beta_0 + X_{i1}\beta_1 + ... + X_{iP}\beta_P, \qquad \text{for } i = 1, \ldots, N \text{ with } P \text{ variables}$$

The regression line with the corresponding parameters can be estimated by minimizing the distance between the regression line and the data points which is called the residual. The residual represents the error term which is formulated as follows:

$$\varepsilon_i = Y_i - E[Y_i], \qquad \text{for } i = 1, \ldots, N$$

6.5.2 Least squared estimator

The parameters in linear regression can be estimated with the least squared estimator (LSE). The LSE determines the parameters by minimizing the sum of squared errors (SSE) which is formulated as follows:

$$\begin{aligned} SSE &= \sum_{i=1}^{N} (Y_i - E[Y_i])^2 \\ &= \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{P} X_{ij}\beta_j \right)^2, \text{ where } X_{i0} = 1 \\ &= (Y - X\beta)^T (Y - X\beta) \end{aligned}$$

Note that the normal equations in the last step have been written in the following matrix notations:

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1P} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{N1} & \cdots & X_{NP} \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_P \end{bmatrix}$$

The LSE of a parameter β_k can be determined by differentiating the SSE with respect to β_k and setting the derivative equal to zero. The derivative of the SSE is calculated as follows:

$$\frac{\partial SSE}{\partial \beta_k} = \frac{\partial}{\partial \beta_k} \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{P} X_{ij} \beta_j \right)^2$$

$$= -2 \sum_{i=1}^{N} X_{ik} \left(Y_i - \sum_{j=0}^{P} X_{ij} \beta_j \right), \text{for } k = 0, ..., P$$

With the derivative of the SSE, the LSE for β_k is derived as follows:

$$0 = \sum_{i=1}^{N} X_{ik} \left(Y_i - \sum_{j=0}^{P} X_{ij} \hat{\beta}_j \right), \text{for } k = 0, ..., P$$

$$0 = \begin{bmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{N1} \\ \vdots & \ddots & \vdots \\ X_{1P} & \cdots & X_{NP} \end{bmatrix} \cdot \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix} - \begin{bmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{N1} \\ \vdots & \ddots & \vdots \\ X_{1P} & \cdots & X_{NP} \end{bmatrix} \cdot \begin{bmatrix} 1 & X_{11} & \cdots & X_{1P} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{N1} & \cdots & X_{NP} \end{bmatrix} \cdot \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_P \end{bmatrix}$$

$$\begin{bmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{N1} \\ \vdots & \ddots & \vdots \\ X_{1P} & \cdots & X_{NP} \end{bmatrix} \cdot \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{N1} \\ \vdots & \ddots & \vdots \\ X_{1P} & \cdots & X_{NP} \end{bmatrix} \cdot \begin{bmatrix} 1 & X_{11} & \cdots & X_{1P} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{N1} & \cdots & X_{NP} \end{bmatrix} \cdot \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_P \end{bmatrix}$$

$$\begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_P \end{bmatrix} = \left(\begin{bmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{N1} \\ \vdots & \ddots & \vdots \\ X_{1P} & \cdots & X_{NP} \end{bmatrix} \cdot \begin{bmatrix} 1 & X_{11} & \cdots & X_{1P} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{N1} & \cdots & X_{NP} \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{N1} \\ \vdots & \ddots & \vdots \\ X_{1P} & \cdots & X_{NP} \end{bmatrix} \cdot \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}$$

$$\hat{\beta} = \left(X^T X \right)^{-1} X^T Y$$

Note that the matrix X can only be inverted when it is nonsingular.

6.5.3 Logistic regression

In logistic regression, the target variable is dichotomous where an event is usually denoted as 1 and a non-event as 0. The probability that an event Y occurs will be denoted as $P[Y = 1]$. The logistic regression model is a generalized linear model and generalized linear models have link functions. The link function connects the probability of event with the parameters. The probability $P[Y = 1]$ has a range from 0 to 1 whereas $X\beta$ can return any value. The link function sets the range of $P[Y = 1]$ equal to $X\beta$, which is in general derived as follows:

1. $P[Y = 1]$ lies in the interval [0, 1].
2. $\frac{P[Y = 1]}{1 - P[Y = 1]}$ lies in the interval [0, ∞]
3. $\log\left(\frac{P[Y = 1]}{1 - P[Y = 1]} \right)$ lies in the interval [-∞,∞]

The resulting link function has the same range as $X\beta$. The link function of the logistic regression model is also called the logit function which is defined as follows:

$$\log\left(\frac{P[Y_i = 1]}{1 - P[Y_i = 1]}\right) = X_i\beta, \quad \text{where } i = 1, \ldots, N$$

Note that the relationship between the probability of event and the parameters is non-linear whereas the relationship between the parameters and the logit function is linear.

When the logit function is rewritten then the probability of event can be formulated as follows:

$$P[Y_i = 1] = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}} = \frac{1}{1 + e^{-X_i\beta}}, \quad \text{where } i = 1, \ldots, N$$

In this case, the target variable is modelled as follows:

$$Y_i \sim Bernoulli\left(\frac{1}{1 + e^{-X_i\beta}}\right), \quad \text{where } i = 1, \ldots, N$$

For positive values of the explanatory variable, when the parameter estimate is positive then this indicates an increasing probability of event. But if the parameter estimate is negative then this indicates a decreasing probability.

The relationship between the target variable and the probability of event can be illustrated with a logistic curve, see figure 6.5 for an example.

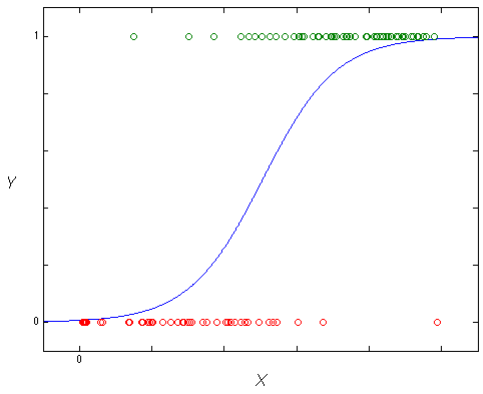

Figure 6.5 Logistic regression curve

The least squared estimators of the parameters in logistic regression can not be explicitly formulated unlike the ones in linear regression, because the relationship between the target variable and the parameters is non-linear. When the SSE in non-linear regression is differentiated with respect to β then this results in a higher order equation which is difficult to solve. The LSE for the parameters in non-linear regression can only be estimated with an iterative algorithm.

6.5.4 Maximum likelihood estimator

The parameters in the generalized linear model can be estimated with the maximum likelihood estimator (MLE). The MLE estimates the parameters by maximizing the log-likelihood function. In logistic regression, the target variable has a Bernoulli distribution where the likelihood function L with the probability of event p is defined as follows:

$$L(\beta) = \prod_{i=1}^{N} p(\beta)^{Y_i} \left(1 - p(\beta)\right)^{(1-Y_i)}$$

The product sign can be changed into a summation by taking the logarithm of the likelihood function, which makes it easier to differentiate to β. This results in the log-likelihood function LL:

$$LL(\beta) = \sum_{i=1}^{N} Y_i \ln\left(p(\beta)\right) + \left(1 - Y_i\right)\ln\left(1 - p(\beta)\right)$$

The MLE can be determined by taking the partial derivatives of LL with respect to β and setting it equal to zero.

6.5.5 Newton-Raphson

The first derivative of LL is called the Fisher's score function which will be denoted as:

$$u(\beta) = \frac{\partial LL(\beta)}{\partial \beta}, \qquad \text{for } N \text{ observations}$$

The matrix of second derivatives of LL is called the Hessian matrix which will be denoted as:

$$H(\beta) = \frac{\partial u(\beta)}{\partial \beta}$$

The Newton-Raphson method approximates the derivative of the log-likelihood function with the first order Taylor series with a set of initial values a for the parameters:

$$u(\beta) \approx u(a) + (\beta - a)H(a)$$

The MLE for β is determined by setting the derivative of the log-likelihood function equal to zero:

$$u(a) + \left(\hat{\beta} - a\right)H(a) = 0$$

$$\hat{\beta} = a - u(a)H(a)^{-1}$$

The parameter estimates can be improved by recomputing the new $\hat{\beta}$ where a is replaced by the old $\hat{\beta}$.

The Newton-Raphson algorithm performs the following steps:

1. Choose a set of initial values for the vector of parameters $\beta^{(i)}$ where iteration $i = 1$.
2. Compute: $\beta^{(i+1)} = \beta^{(i)} - u\left(\beta^{(i)}\right)H\left(\beta^{(i)}\right)^{-1}$.
3. If the difference between $\beta^{(i)}$ and $\beta^{(i+1)}$ exceeds the convergence level, then go to step 2 and raise i by 1. Stop otherwise.

The number of iterative steps in the Newton-Raphson method depends on the form of the log-likelihood function, on the starting values of the parameters and on the convergence level. If the log-likelihood function is close to quadratic or if the starting values is close to the MLE or if the convergence level is large, then the Newton-Raphson procedure will converge quickly.

6.5.6 Fisher's scoring

Fisher's score function has a zero mean at the true values of the parameters:

$$E\left[u(\beta)\right] = 0$$

The corresponding covariance matrix is given by Fisher's information matrix:

$$\begin{aligned} I(\beta) &= Var\left[u(\beta)\right] \\ &= E\left[u(\beta)u^T(\beta)\right] - E\left[u(\beta)\right]E\left[u(\beta)\right]^T \\ &= E\left[u(\beta)u^T(\beta)\right], \text{ since } E\left[u(\beta)\right] = 0 \end{aligned}$$

Under mild regularity conditions, Fisher's information matrix can also be obtained as follows:

$$I(\beta) = -E\left[H(\beta)\right].$$

Fisher's information matrix can be estimated when replacing β in the right hand side of the equation above by the maximum likelihood estimator $\hat{\beta}$.

The Fisher's scoring method replaces minus the Hessian matrix in the Newton-Raphson method by Fisher's information matrix which results in the following approximation of $\hat{\beta}$:

$$\beta^{(i+1)} = \beta^{(i)} + u\left(\beta^{(i)}\right)I\left(\beta^{(i)}\right)^{-1}, \text{ for iteration } i$$

The Fisher's scoring and the Newton-Raphson's method calculate the same parameter estimates, except the corresponding estimated covariance matrix may differ slightly. This is due to the fact that the Fisher's scoring method is based in the expected information matrix while the Newton-Rahpson method is based on the observed information matrix. For the logistic regression model, the observed and the expected information matrices are identical, which results in identical estimated covariance matrices for both algorithms.

6.6 Variable selection

The objective in building the regression model is to maximize the goodness-of-fit of the data sample against a minimal number of explanatory variables. But the number of variables rises with the goodness-of-fit. Therefore, a trade-off needs to be made between the goodness-of-fit and the number of variables.

When all the available explanatory variables are included in the model, then this results in the full model which probably includes variables that have an insignificant relationship with a target variable. The significance of the variable can be indicated by performing a statistical test. The variable selection methods exclude or include variables in the model by means of tests of independence.

6.6.1 Wald test

The Wald test examines whether the explanatory variable, which is included in the logistic regression model, has a significant relationship with the target variable. The null hypothesis H_0, the alternative hypothesis H_1 and the Wald statistic W are defined as follows:

$$\begin{aligned} &H_0 : \beta_j = 0 \\ &H_1 : \beta_j \neq 0 \end{aligned}, \quad \text{where } 0 \leq j \leq P+1$$

$$W = \frac{\hat{\beta}_j^{\,2}}{\hat{Var}\left[\hat{\beta}_j\right]},$$

with MLE $\hat{\beta}_j$

$$\hat{Var}[X] = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$$

The Wald statistic has a chi-square distribution with one degree of freedom under the null hypothesis. When the corresponding p-value is greater than the significance level α, then the Wald test fails to reject the null hypothesis and the variable is removed from the model. Failure of rejecting the null hypothesis indicates that removing the variable from the model will not substantially harm the fit of the model, since an explanatory variable with a parameter that is very small relative to its standard error is in general not helpful in predicting the target variable.

6.6.2 Score test

The Wald test and the score test (also known as Lagrange Multiplier test) both examine nested models, but the score test is slightly different. When applying the score test, then it is not necessary to estimate the parameter of interest β_{P+1}, since the score statistic does not include the parameter. This means that the score test requires less computational effort than the Wald test. The score test can then be used to examine whether adding the explanatory variable to the model will result in a significant improvement in model fit. The null and alternative hypotheses and the score statistic are formulated as follows:

$$\begin{aligned} &H_0 : \beta_j = 0 \\ &H_1 : \beta_j \neq 0 \end{aligned}, \quad \text{where } 0 \leq j \leq P+1$$

$$S = u^T\left(\hat{\beta}_{H0}\right) I^{-1}\left(\hat{\beta}_{H0}\right) u\left(\hat{\beta}_{H0}\right),$$

with MLE $\hat{\beta}_{H0} = \left(\hat{\beta}_0, \ldots, \hat{\beta}_{P+1}\right)$ where $\hat{\beta}_j = 0$.

The score statistic is based on the slope of the likelihood function which is used to estimate the improvement in model fit if the variable is added to the model. Under the null hypothesis, the score statistic has an asymptotically chi-square distribution with one degree of freedom. When the corresponding *p*-value is equal to or lower than the significance level α, then the null hypothesis of the score test is rejected and the variable is added to the model.

Figure 6.6 shows a graphical illustration of the relation between the score and Wald tests, where the parameter value is on the horizontal axis and the loglikelihood *LL* on the vertical axis. The Wald test is based on the indicated horizontal distance, whereas the score test on the indicated gradient.

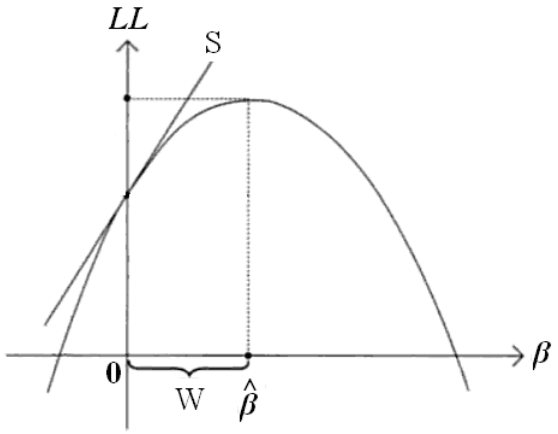

Figure 6.6 Comparison of Wald and score test

6.6.3 Selection methods

There are several variable selection methods:

- Backward
- Forward
- Stepwise
- Best subset

The score test is suitable for including significant variables to the model, because the score approach essentially starts at the null hypothesis and examines whether movement towards the alternative hypothesis will be an improvement. The Wald approach on the other hand starts at the alternative hypothesis and considers movement towards the null hypothesis, which makes the Wald test more suitable for removing insignificant variables from the model. Therefore, the backward selection approach applies the Wald test to remove variables, the forward selection approach applies the score test to include variables and the stepwise selection approach applies both the score and Wald test.

The backward selection method performs the following steps:

1. Estimate the parameters for the full model.
2. Compute the Wald statistic and the corresponding *p*-value for each individual parameter in the current model.
3. Perform the Wald test: if the highest *p*-value > α then do not reject H_0, remove the corresponding variable from the model and go to step 2. Stop when otherwise.

When the number of explanatory variables is too large or when the sample size is too small, then the backward selection method is usually not suitable due to complete or quasi-complete separation of the target values in the first iteration step. In case of complete separation, there is one explanatory variable or a linear combination of variables that perfectly predicts the target value. Then it is not possible to compute the maximum likelihood estimates of the parameters because the slope of the logistic function would be infinite.

Figure 6.7 shows an example of complete separation with one explanatory variable X, where all observations of X below 0.5 are classified as $Y = 0$ and all observations of X above 0.5 are classified as $Y = 1$. This results in an infinite maximum likelihood estimate of β. In case of quasi-complete separation, the target values overlap or are tied at a single value of an explanatory variable. Then the parameter estimates or the corresponding standard errors are extremely large.

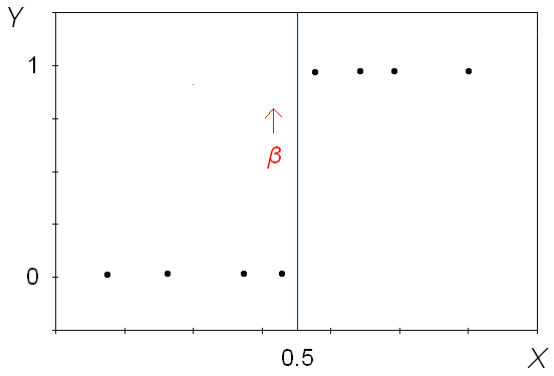

Figure 6.7 Complete separation of data points

The forward selection method performs the following steps:

1. Estimate the intercept for the model with 0 variables.
2. Compute the score statistic and the corresponding p-value for each individual parameter not in the current model.
3. Perform the score test: if the lowest p-value $\leq \alpha$ then reject H_0, add the corresponding variable to the model and go to step 2. Stop when otherwise.

The forward selection method will not return a complete or quasi-complete separation of the target values in the first iteration step.

The stepwise selection method performs the following steps:

1. Estimate the intercept for the model with 0 variables.
2. Compute the score statistic and the corresponding p-value for each individual parameter not in the current model.
3. Perform the score test: if the lowest p-value $\leq \alpha$ then reject H_0 and add the corresponding variable to the model.
4. Compute the Wald statistic and the corresponding p-value for each individual parameter in the current model.
5. Perform the Wald test: if the highest p-value $> \alpha$ then do not reject H_0, remove the corresponding variable from the model and go to step 4. Otherwise, go to step 2. Stop when there are no variables to be added to the model or when the recently added variable has been removed in the current step.

When a variable has been removed from or added to the model in respectively the backward or forward selection method, then the variable remains excluded or included. But the stepwise selection method is a combination of the forward and backward selection method where the initial model has zero variables. This means that the added variable will not necessarily remain in the model.

The stepwise selection method can also take suppression effects into account by performing an elimination step after adding each variable. A suppression effect is present when a suppressor variable has a weak (or strong) effect by itself but takes on a stronger (or weaker) effect after combining it with another predictor. The forward selection method adds one variable at the time, which does not make it possible to examine a weakening effect of the suppressor variable when adding another variable. The backward selection method on the other hand may remove a suppressor variable with weak individual effect, which is only strong when combined.

In the best subset selection method, all possible models are built that contain every combination of variables where the best subset is selected with the highest likelihood score statistic. The computational complexity of the subset selection method rises rapidly with the number of variables. This selection method is not available for models with categorical variables in SAS.

6.7 Dummy variables

The parameters β assign an equal weight for all the values of the corresponding explanatory variable X. It is possible that the relationship between the logit function and the explanatory variable is non-linear and non-monotonic for the continuous explanatory variable. This problem can be solved by applying a different non-linear function instead of the logit function or by creating dummy variables.

Dummy variables are multiple dichotomous variables which are created for each explanatory variable where ‘1’ usually means that the value belongs in the category and ‘0’ when otherwise. When dummy variables are created, then a separate regression coefficient will be estimated for each category of the explanatory variable. When the dataset contains missing values, then the latter one is more suitable.

6.7.1 Bin and coarse classification

The classification of the explanatory variable can be executed by means of the contingency table where the relative frequency of the target value is examined for each pre-defined category of the explanatory variable. The classification procedure is in general divided in two steps:

1. Bin classification: fine classify the explanatory variables in groups which are small enough to capture the distinctive relationship with the target variable and which are also large enough to reduce noise in the data and be representative for the entire population.
2. Coarse classification: combine similar bin classes with respect to the relationship with the target variable.

The bin classification is considered as exploratory data analysis and also as a first step in developing a statistical model. The data errors can also be discovered during the bin classification. It is in general not clear how to define the bin classes, which is why a considerable amount of time should be spent on alternative bin classes. In case of continuous variables, each bin class usually contains approximately a fixed percentage of the data where the borders of the bin classes are conveniently rounded values. The bin classes can also be formed in such a way that it returns a monotonic relationship when this has been expected.

The dummy variables can be created for all types of explanatory variables. If the explanatory variable is continuous then only the neighbouring groups are combined in the coarse classification step. But if the explanatory variable is categorical then the groups with a similar relationship with the target variable or the groups with similar characteristics are combined. When the groups with similar characteristics are combined, then this requires domain knowledge.

The advantage of the bin and coarse classification is that the relationship with the target variable in the regression analysis is examined per category of the explanatory variable instead of the whole domain. When the continuous variable is categorized in separate dummy variables, then this makes the regression model more robust against outliers.

The classification of explanatory variables also has disadvantages. The specific values of the variables will be lost in the classification process which results in a loss of information. When the value of the variable is close to the classification border then it is classified in either one class or the other.

The bin and coarse classification of the explanatory variables requires analyst experience, preferences matters and a combination of art and science. There are no specific steps for classifying the variables, only a general guideline. The most suitable approach depends on the context which includes the goal of classification and the model specifications.

6.8 Confidence interval

The confidence interval of the parameter indicates the reliability of the parameter estimate by showing the estimated range of values which is likely to include the the true parameter value. The $1 - \alpha$ confidence interval for parameter β is formulated as follows:

$$P\left[L(X) \leq \beta \leq U(X)\right] \geq 1 - \alpha ,$$

with lower limit $L(X)$ and upper limit $U(X)$ of the confidence interval as functions of observations X

The confidence interval should be considered as follows: suppose that n random samples are obtained in exactly the same way as the observations and the confidence interval is computed n times for each sample, then it is expected that at least approximately $n(1 - \alpha)$ of the intervals contain the true parameter value. This has been illustrated in figure 6.8, where there are $n = 48$ computed 95% confidence intervals, of which three intervals do not contain the true parameter value. In reality, it is not possible to repeat the computation of the confidence interval. Therefore, the computed confidence interval may not contain the true parameter value, but this can not be verified.

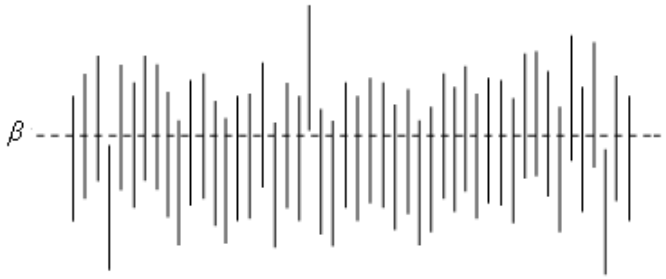

Figure 6.8 Confidence intervals for β based on random samples

When α is small, then the confidence interval is larger. In this case, the confidence interval is more likely to include the true parameter value, but it will return less information about the parameter. Therefore, a trade-off needs to be made when determining α. A commonly chosen value for α is 0.05.

In SAS, there are two methods for computing the confidence interval for the parameters which are the Wald confidence interval and the likelihood ratio-based confidence interval. These methods will be described in the following sections.

6.8.1 Wald confidence interval

The Wald confidence interval is also called the normal confidence interval since it is based on the asymptotic normality of the parameter estimators. The $(1 - \alpha)$ Wald confidence interval for parameter β_j is formulated as follows:

$$\hat{\beta}_j \pm z_{1-\alpha/2}\hat{\sigma}_j,$$

where

$\hat{\beta}_j$ = MLE of parameter β_j where $j = 0, \ldots, P$

z_p = p^{th} percentile of standard normal distribution

$\hat{\sigma}_j$ = estimated standard error of $\hat{\beta}_j$

6.8.2 Profile likelihood confidence interval

The profile likelihood confidence interval is also known as the likelihood ratio-based confidence interval. It is based on the profile likelihood function which is defined as follows for β_j:

$$LL_j^*(\beta_j) = \max_{\beta \in B_j} LL(\beta),$$

where

$LL(\beta)$ = log likelihood function for $\beta = (\beta_0, \ldots, \beta_P)$

$B_j = \beta$ with the j^{th} element fixed at β_j

If $\hat{\beta}$ is the MLE for β and β_j is the true parameter value then $2\left(LL\left(\hat{\beta}\right) - LL_j^*\left(\beta_j\right)\right)$ has a limiting chi-square distribution with one degree of freedom. The $(1 - \alpha)$ profile likelihood confidence interval for β_j is defined as follows:

$$LL_j^*\left(\beta_j\right) \geq LL\left(\hat{\beta}\right) - 0.5\chi^2_{1-\alpha,1},$$

where

$\chi^2_{p,q}$ = p^{th} percentile of chi-square distribution with q degrees of freedom

The lower and upper limit of the profile likelihood confidence interval in SAS are approximated by an iterative procedure which is more time-consuming than the Wald confidence interval. But the profile likelihood confidence interval is more accurate, especially when the sample size is small.

6.9 Lift curve

In binary classification, the outcome is usually labelled as either positive or negative. The confusion matrix in table 6.4 shows the four types of outcomes. The true positive rate and true negative rate are also called respectively sensitivity and specificity. Note that the false positives are considered to be Type I errors and the false negatives Type II errors.

		ACTUAL CATEGORY	
		Event	**Non-event**
PREDICTED CATEGORY	**Event**	True positive (TP)	False positive (FP)
	Non-event	False negative (FN)	True negative (TN)
	Total	Positives (P)	Negatives (N)

Table 6.4 Confusion matrix for binary classification

The lift curve (also called gains chart or Lorenz curve) is a graphical technique which is used to evaluate the quality of classification models like the CHAID tree and the logistic regression model at a certain cut-off point. The lift curve shows how many of the events are correctly classified when a part of the data sample is selected, where the data sample is ranked by the predicted probability of the event.

When the classification model returns the estimated probability of an event for each observation, then the lift curve is constructed by taking the following steps:

1. Order the N observations in such a way that the estimated probability is descending.
2. For each observation n, compute the cumulative number of observed positives relative to the total number of positives P in order to calculate the TP rate:

$$\text{TP rate}(n) = \frac{\sum_{i=1}^{n} Y_i}{P}, \text{ for observation } n = 1, \ldots, N$$

where

$$Y_i = \begin{cases} 1, & \text{if observation } i \text{ is labelled as positive} \\ 0, & \text{if observation } i \text{ is labelled as negative} \end{cases}$$

$$P = \sum_{i=1}^{N} Y_i$$

3. Plot the TP rate on the vertical axis and the percentage of the total number of observations on the horizontal axis, see figure 6.9 for an example of the lift curve. The curved line represent the lift curve and the linear line is the random classifier. The greater the area between the linear line and the lift curve, the better the model. Usually, the set of plotted datapoints is not too large in order to return a smooth curve.

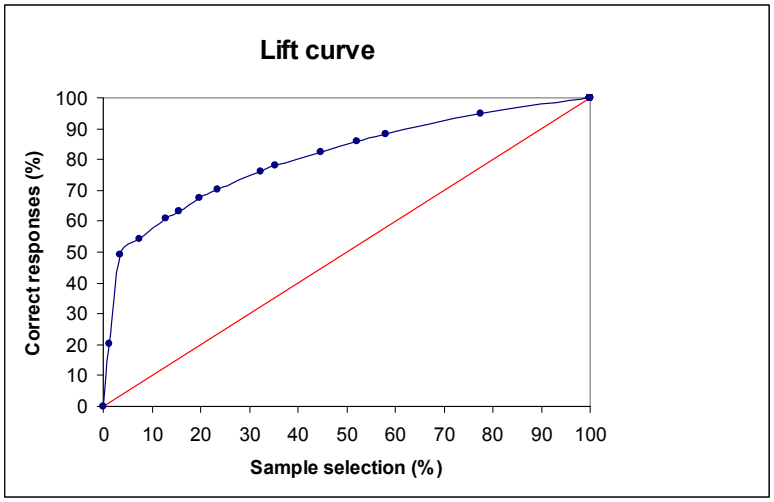

Figure 6.9 Lift curve

Note that the lift curve depends on the ratio of positives to negatives in the sample selection.

6.10 ROC curve

The receiver operating characteristic (ROC) curve is similar to the lift curve, except the horizontal axis is different. The ROC curve is plotted by defining the (x, y) coordinates on the horizontal x-axis and vertical y-axis as follows:

$$x = \frac{FP}{N}$$
$$y = \frac{TP}{P}$$

The ROC curve makes a relative trade-off between the benefit of true positives against the cost of false positives. An example of the ROC curve is shown in figure 6.10 where the curved line represents the ROC curve and the linear line is the random classifier. The greater the area between the linear line and the ROC curve, the better the model. The ROC curve resembles the lift curve when the FP rate or the TP rate is small.

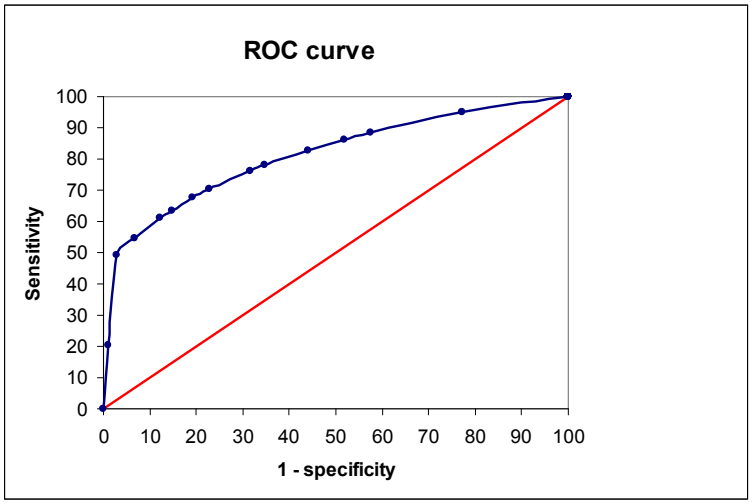

Figure 6.10 ROC curve

An advantage of the ROC curve is that it is independent of the ratio of positives to negatives. This makes the ROC curve suitable for comparing different classification models when the ratio of positives to negatives can vary over time.

6.10.1 Area under curve

One way to measure the ROC curve is the *c*-value which is defined as follows:

$$c\text{-value} = \frac{\text{area under ROC curve}}{\text{total area}}.$$

The random classifier has a *c*-value of 0.5 and can be compared with tossing a coin whereas the perfect classifier has a *c*-value of 1. Therefore, the lower and upper bounds of the *c*-value are respectively 0.5 and 1. The higher the *c*-value, the better the model. The *c*-value of the ROC curve in figure 6.8 is 0.81. The overall guideline for interpreting the *c*-value is as follows:

- 0.5 – 0.6: fail
- 0.6 – 0.7: poor
- 0.7 – 0.8: fair
- 0.8 – 0.9: good
- 0.9 – 1: excellent

When the data sample is representative for the whole population, then the *c*-value can also be considered as the probability that the model correctly classifies the event.

6.11 Hosmer-Lemeshow test

The goodness of fit of the model can be examined with the Hosmer-Lemeshow test. The Hosmer-Lemeshow test indicates how well the predicted probabilities agree with the actual risk. The Hosmer-Lemeshow test is performed as follows:

1. Rank the observations from low to high probability of event.
2. Split the observations in M evenly distributed groups.
3. The null hypothesis and the Hosmer-Lemeshow statistic HL is defined as follows:

   H_0: good fit

$$HL = \sum_{i=1}^{M} \frac{(q_i - r_i)^2}{r_i},$$

   where

   q_i = actual number of events in group i
   r_i = average predicted probability × number of observations in group i

Under the null hypothesis, HL has a chi-square distribution with $(M - 2)$ degrees of freedom. If the p-value $\leq \alpha$, then the null hypothesis is rejected and the model has a lack of fit.

6.12 Cut-off point with optimal classification accuracy

The logistic regression model returns for each case the estimated probability that the value of the target variable is positive (or negative). When the estimated probability is equal to or greater than the cut-off point, then the case will be classified as positive (or negative). For an estimation less than the cut-off point, the case will be classified as negative (or positive).

The cut-off point with the optimal classification accuracy maximizes the TP rate against a minimal FP rate. This can be determined by means of the ROC curve. The cut-off point with optimal classification accuracy corresponds to the point on the ROC curve where the difference between the TP rate and FP rate is maximized:

$$\max(diff) = \max\left\{\frac{TP}{P} - \frac{FP}{N}\right\}$$

Therefore, the cut-off point with optimal classification accuracy is the estimated probability where the difference is maximal. In practice, the cut-off point is not determined by the optimal classification accuracy but by the risk exposure level of the bank. When the risk exposure level of the bank is 5%, then the cut-off point is determined by the estimated probability where the Type I error is 5%.

7 Correlation matrix and contingency table

[Classified]

8 Cluster analysis

[Classified]

9 CHAID tree

[Classified]

10 Logistic regression

[Classified]

11 Conclusion and recommendation

[Classified]

References

Abdi, H. (2007), '*Bonferroni and Sidak corrections for multiple comparisons'*. In: Neil Salkind: Encyclopedia of Measurement and Statistics, Thousand Oaks, CA: Sage.

Anderson, R. (2007), '*The credit scoring toolkit, theory and practice for retail credit risk management and decision automation'*, Oxford University Press, UK.

Eherler, D. and Lehmann, T. (2001), '*Responder profiling with CHAID and dependcy analysis'*. Friedrich-Schiller-Universität, Jena.

Gunst, M (2005), '*Statistical models*'. Lecture notes, Vrije Universiteit, Amsterdam.

Guyon, I. and Elisseeff, A. (2003), *'An introduction to variable and feature selection'*, Journal of Machine Learning Research 3, pp. 1157-1182, Cambridge.

Hosmer, D. and Lemeshow, S. (1989), *'Applied Logistic Regression'*, John Wiley & Sons, New York.

Jung, K. and Thomas, L (2008), *'A note on coarse classifying in acceptance scorecards'*, Journal of the Operational Research Society, Volume 59, Number 5, pp. 714-718(5), Hampshire.

Menard, S. (1995), *'Applied Logistic Regression Analysis'*, Sage Publications, California.

Oosterhoff, J. and Van der Vaart, A.W. (2003) *'Algemene Statistiek'*. Lecture notes, Vrije Universiteit, Amsterdam.

Pampel, F. (2000), *'Logistic Regression: A Primer'*, Sage Publications, California.

Tan, P. N, Steinback, M. and Kumar, V. (2005), '*Introduction to data mining'*, Addison Wesley, pp. 487-568, London.

Vuk, M. and Curk, T. (2006) '*ROC curve, lift chart and calibration plot*', Metodološki zveski, vol. 3, no. 1, pp. 89-108, Slovenia.

Appendix A: client overview

[Classified]

Appendix B: application & contract overview

[Classified]

Appendix C: accepted application overview

[Classified]

Appendix D: delinquency & transaction overview

[Classified]

Appendix E: summary statistics

[Classified]

Appendix F: ranked correlation matrix

[Classified]

Appendix G: contingency tables

[Classified]

Appendix H: cluster analysis of applications

[Classified]

Appendix I: SAS code

[Classified]