Örebro University

Örebro University School of Business

Master in Applied Statistics

Thomas Laitila

Sune Karlsson

May, 2014

CREDIT SCORING MODEL APPLICATIONS:

TESTING MULTINOMIAL TARGETS

Gabriela De Rossi Ayres (85/09/11)

Wei Wei (87/12/31)

Acknowledges

We would like to thank all people who helped us, direct or indirectly, to complete our master thesis, by giving advices, supporting and motivating us or just cheering for us.

In special to our parents, Carlos and Rafaela, Wei and Zheng, who supported us since when studying master abroad was just a dream.

To our supervisor, Thomas Laitila, who was always willing to, patiently, guide us and still give us the freedom to work with our ideas and opinions.

To Sune Karlsson, our examiner who kindly accepted to grade our work.

To all the faculty members and colleagues from Örebro University that in some way contributed to our study experience and made our days happier and more pleasant.

And finally, we would like to thank each other, for the commitment and fellowship that just got better by writing this work together.

Abstract

For the construction and evaluation of credit scoring, one commonly used way to classify applicants is using the logistic regression model with binary target. However, other statistical models based on multinomial targets can be considered, such as ordered logistic regression model, generalized ordered regression model and partial proportional odds model. These models are tested in real data and comparisons are made to analyze the most appropriated option aiming for different proposes.

Keywords: credit scoring, neutral target, logistic regression model, ordered logistic regression model, generalized ordered regression model, partial proportional odds model, model comparison.

Contents

1 - INTRODUCTION 1

1.1 - Credit Risk Evaluation 2
1.2 - Micro Loan 4
1.3 - The process 5

2 - THE DATASET 7

3 - MODEL COMPARISON: EFFECT OF EXCLUDING NON-RANDOM PART OF THE SAMPLE .. 17

3.1 – Logistic Regression: A Common Approach in the Financial Market 17
3.2 – Theoretical Foundation 20
3.2.1 - Logistics regression 20
3.2.2 - Ordered logistic regression 22
3.3 – Model comparison with data exclusion 26
3.4 – Case Illustration 29

4 - NEW MODEL MOTIVATION 36

4.1 - Generalized Ordered Logit Model 39
4.2 - Partial Proportional Odds Model 40

5 - APPLICATION AND RESULTS 41

6 - CONCLUSIONS 56

REFERENCES 58

APPENDIX 61

Chapter 1

Introduction

The granting of credit plays a fundamental role in the economy of a country. Institutions that extend credit in exchange for a gain on borrowed capital adopt procedures to decide whether or not to lend money to an applicant. The goal is to reduce erroneous approvals and rejections of claims profitable. Though the common way of granting the credit in current market is more operable and understandable for business running, it cannot avoid facing the problem in model evaluation and prediction. The purpose of this paper is, especially from statistical perspective, to derive a suitable model based on a multinomial target, meanwhile keeping its viability.

For market purpose, the common application is built based on binary target which simply define customers into *"bad"* and *"good"*. However it would be interesting to identify customers' behavior in a more detailed way, by adding one more classification category, *"neutral",* which results in a multinomial target. The first part of this paper will compare binomial and multinomial targets using theoretical proof and a practical case application. Furthermore, different models for the multinomial target will be proposed, constructed and compared in second part of paper. Combined with market case, the paper will suggest the most appropriate and practical model from both market and statistical point of view, based on the development sample.

1.1 - Credit Risk Evaluation

Credit Risk Evaluation is one of the main areas in a well-structured financial institution. Statisticians are highly required to control risks and find out new opportunities. It is in this promising environment that the results of this study will be obtained.

According to the Bank of Mauritius (2003), credit processing is the stage when all required information on credit is gathered and applications are screened. A pre-qualification screening criteria is set, which would act as a guide. For instance, the criteria may include rejecting applications from blacklisted customers or other cluster of applications/customers that would be processed and rejected later. Moreover, this stage is important to avoid fraudulent activities or activities that are against the law what could damage the institution's reputation.

The next stage is to check customer's ability to meet his payment obligations (Ibid). A list of policy rules is established by the Risk Area in order to decline applications in which profiles are not of interest. At this stage, external and internal information is needed.

External information can be obtained from credit bureaus, which are corporations that collect information from different sources and provide consumer credit information on individual's borrowing and bill-paying habits (Sullivan & Sheffrin, 2003, p. 512). It comes typically from creditors, lenders, utilities, debt collection agencies and the courts that a consumer has had a relationship with (Ibid). The availability of external information depends on the legislation of each country. Some countries have a strict policy and no personal information can be shared. On the other hand, there are countries where any personal information can be accessed, positive or negative information.

Examples of external information are the following: registered address, yearly income, income tax, bankruptcy, trustee, remarks (when customers have paid after the due date and institutions

report it to a credit association), bad debts (a remark becomes a bad debt when the process go to court), and so forth. Examples of internal information that may be relevant at this stage can be whether customer has already an open loan, whether he/she still has an unused monthly limit to borrow or whether he/she is in "cooling period" due to late payment in previous loans.

The last stage is to classify the potential applications according to their risk and make a decision. It is here where the credit score will take place. A high quality score, developed with the appropriated technique, contains strong variables able to efficiently explain the default and applies with an appropriate strategy that can be the key for increasing profits.

Lenders use credit scores to determine who qualifies for a loan, determine the interest rates, set the loan limit and mitigate losses due to bad debts. Credit scores enable the rapid decision-making and count with the probability theory which is more reliable than the common sense opinion from a loan handler.

1.2 - Micro Loan

Data used in this paper come from a lending company that will be codenamed as “WG Money”. WG Money is considered a micro loan institution, which provides small loans to be paid back in short periods, in a fast and easy way. Micro loans are obtained increasingly through the internet or mobile text messages/apps instead of store outlets. Customers do not need to send a lot of information, usually just the identification number, amount to be borrowed, term, address and bank account. The application is processed in less than ten minutes and, in case it is approved, customer can have the money transferred immediately to his/her account.

The main advantage of this business is how fast and accessible it is, therefore the need of a short form. The success comes by obtaining information about customers from other sources. This will guide the approval decision and indicates how likely is to have the payment back and, thus, the limit amount to be lent.

Data confidentiality is very strict since having information about customers is the main way to make accurate decisions and, consequently, business a success.

1.3 - The process

During the past eight years, this micro loan company has increased its market share. The need for a well-structured risk area brought investments in appropriate software and high-qualified employees specialized in scorecards.

The intention is to test different techniques to develop a new scorecard model to returning customers, leading to a final model which is capable to distinguish customers according to their risk to take credit.

The application process is simple: customers send an application by mobile phone or by the website. The application is processed by the "loan handling system" and all internal and external information available is collected and shown in the screen to the loan handler, including the policy rules and scorecard. Loan handlers send immediately the decision to the customer's e-mail or mobile phone. If the application is approved, customer sends the confirmation to take the loan and, on the other side, the loan handler transfers the money to customer's account. The loan amount goes from 50 to 600 euros to be paid in one single installment after 30 days.

After the loan is granted, the following step is to receive the payment from customers. The figure below shows the scheme for the actions taken by the collection area when a payment delay happens:

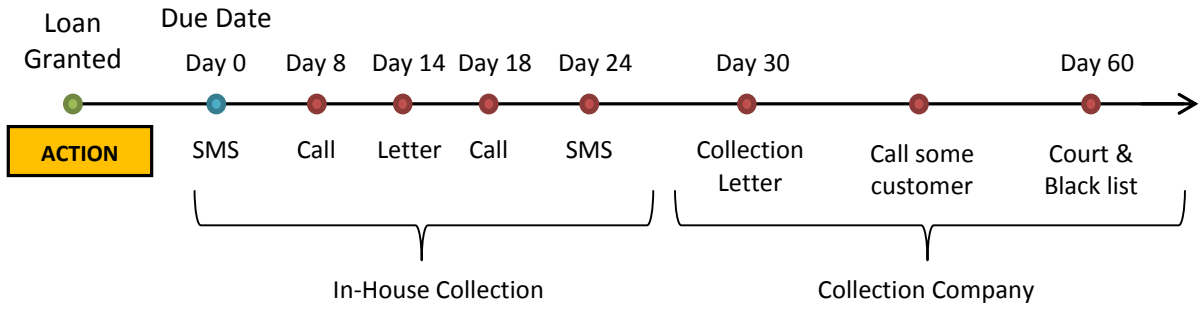

Figure 1.1 : **Standard process for receiving payment from customers**

The first collection stage is in-house. Reminder letters, SMS and phone calls are channels to communicate with customers to remind them about the payment. From 30 to 60 days from the due date, it becomes responsibility of an outsourced company. At the end of this period, customers are reported to a loan association and this remark will become public. Moreover, the customer will be sent to court as a last try to receive the money back.

The graphic below shows the relation between the accumulated payment rate and the number of days after the due date when the customer paid back his loan. Around 52% of customers pay back the loan until the due date. During the in-house collection period, extra 39% pay back, during the outsourced collection period more 4% and, hence, before the lawsuit, 95% of the customers in this period have paid back their loans.

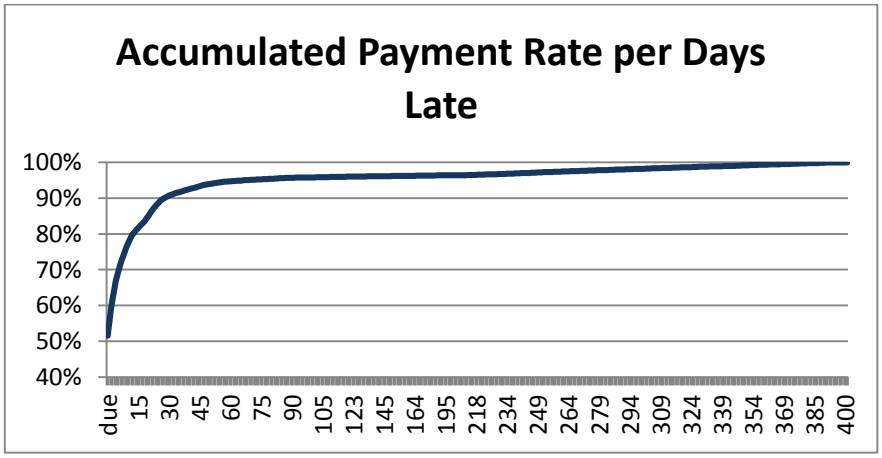

Figure 1.2: **Accumulated payment rate per days late**

Chapter 2

The dataset

Dataset consists of all applications received in the period of January to June 2011, summing up 29,873 observations. All these applications are from former customers, which make it possible to use their behavior information based on previous loans.

From all applications, approximately 84% are approved. From all approved, 92% were paid out to the customers and are, then, called 'loans'. Models will be based on loans data, which performance is known. It means that it is known how late customers have paid back their loans. From that, targets – or dependent variables - were defined: Target1 classifies customers in the categories *"good", "bad"* and *"neutral",* according to the payment behavior: number of days from the payment date to the due date. Target2 classifies in *"good"* and *"bad"*. Target3 classifies customers as Target1. Each target is going to be used to develop a different model, different by the technique and/or sample. More details about the targets will be presented in Chapter 3.

The proportion for Target1 and Target3 is about 6% of *"bad"* customers, 10% of 'neutral' and the remaining as *"good"*. Target2 has 6% of *"bad"* customers and 94% of *"good"*. Both Target1 and Target2 will be applied as binary response: *"neutral"* customers according to Target1 are not considered for developing the model. This will be better explained in Chapter 3. The above may cause bias and this will be evaluated. Results will be compared with the model fit with Target2.

All available independent variables were collected to possibly explain the default: demographic information like age, gender, zip code, income, marital status; behavior information like how many loans were granted to this customer, for how long time he/she has been a customer, how

late he/she paid back previous loans, etc. They were pre-analyzed to check the relation with the dependent variable. Just those which presented strong relation with the dependent variable were selected to the following steps of the development.

The names of variables are encoded because of confidentiality reasons, but it doesn't have negative effect to the results since the interest is to compare the models in how well they distinguish "good"/"bad" customers, and not the model itself.

The variables are taken and categorized into a relatively small number of groups. The final categorization is given by collapsing some of these groups in a way that each will have enough data to fit the model (Kočenda & Vojtek, 2009). It is also includes the continuous variables, which is a common approach in credit scoring. *"For continuous characteristics, the reason is that credit scoring seeks to predict risk rather than to explain it, and so one would prefer to end up with a system in which the risk is nonlinear in the continuous variable if that is a better prediction."* (Lyn, Edelman, & Crook, 2002).

The groups were created based on the following criteria:

- Similar bad rate within groups;
- Highest difference in bad rate as possible between groups;
- Frequency: groups should have at least 5% of the observations to be considered consistent.

(q-1) dummy variables are created, where *q* is the number of groups. One group will be set as "reference cell", chosen the one which bad rate is closest to the average bad rate (6%). When fitting the model, all the dummy variables will be tested.

Below, descriptive statistics and figures are presented about the variables to be tested in the models:

Variable V1

V1 is a numeric discrete ordinal variable, with possible outcomes from 18 to 99 and average is equal to 33.5. It expresses demographic characteristics of the customer. *V1* is categorized into four groups, *V1_1*, *V1_2, V1_3* and *V1_4*. The third dummy is taken as the reference cell.

The graph below shows the bad rate and frequency of each *V1* possible outcome followed by the table expressing the characteristics for each group created.

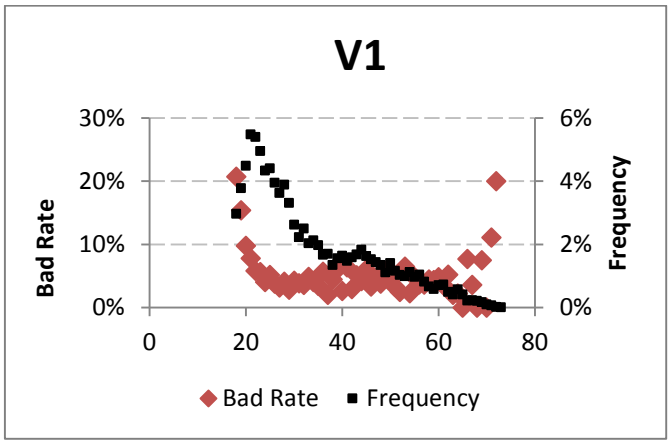

Figure 2.1 : **Bad rate and frequency of explanatory variable V1**

Variable V1						
Categories	Band	# Bad	# Good	# Total	% Total	% Bad
V1_1	18 - 19	278	1286	1564	7%	17.8%
V1_2	20 - 21	201	2110	2311	10%	8.7%
V1_3	22 - 35	479	10501	10980	47%	4.4%
V1_4	36+	352	7959	8311	36%	4.2%
Total		**1310**	**21856**	**23166**	**100%**	**6%**

Table 2.1 : **Characteristics summary of explanatory dummy variable V1**

Variable V2

V2 is a numeric discrete variable that goes from 1 to infinite and which average is equal to 8.5. It is a behavior variable which expresses information from previous loans taken by the customer. *V2* is categorized into seven groups, *V2_1, V2_2, ... , V2_7*. The reference cell for *V2* is *V2_4*.

The graph below shows the frequency and bad rate for *V2*. It shows high concentration of observations and high bad rate for lower possible values of *V2*. The table displays characteristics of the created groups.

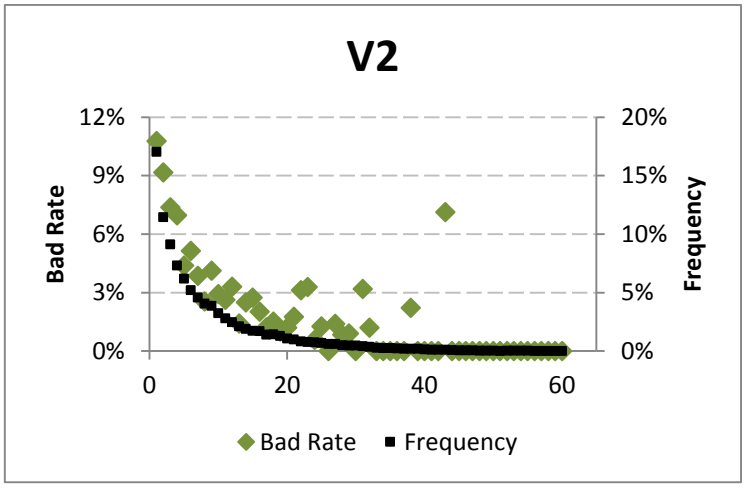

Figure 2.2 : **Bad rate and frequency of explanatory variable V2**

Variable V2

Categories	Band	# Bad	# Good	# Total	% Total	% Bad
V2_1	1	426	3523	3949	17%	10.8%
V2_2	2	243	2408	2651	11%	9.2%
V2_3	3 - 4	274	3531	3805	16%	7.2%
V2_4	5 - 6	125	2517	2642	11%	4.7%
V2_5	7 - 9	102	2791	2893	12%	3.5%
V2_6	10 - 15	87	3220	3307	14%	2.6%
V2_7	16+	53	3866	3919	17%	1.4%
Total		**1310**	**21856**	**23166**	**100%**	**6%**

Table 2.2 : **Characteristics summary of explanatory dummy variable V2**

Variable V3

V3 is a variable of behavior type. It takes values from 0 to infinite as a discrete variable. The concentration of observations is the value 0, with frequency of more than 60%. The average is 0.98. Two groups are created: *V3_1* and *V3_2*, and the first group is chosen to be the reference cell. More details about this variable can be found in the graph and table below:

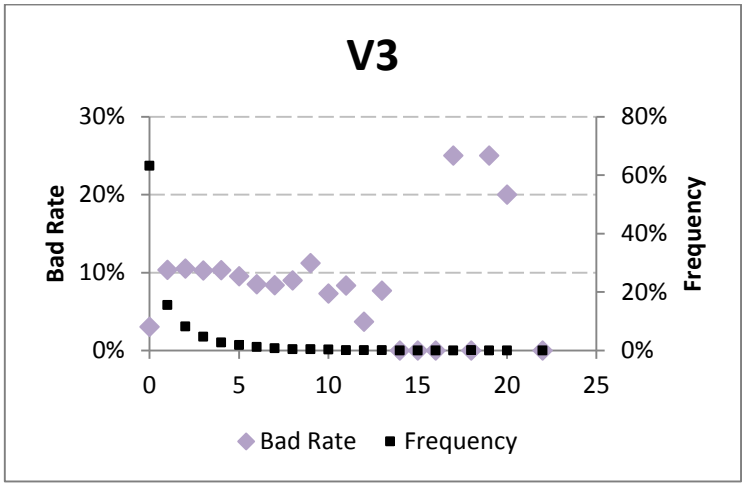

Figure 2.3 : **Bad rate and frequency of explanatory variable V3**

Variable V3

Categories	Band	# Bad	# Good	# Total	% Total	% Bad
V3_1	0	445	14194	14639	63%	3%
V3_2	1+	865	7668	8533	37%	10%
Total		**1310**	**21856**	**23166**	**100%**	**6%**

Table 2.3 : **Characteristics summary of explanatory dummy variable V3**

Variable V4

It is a discrete variable taking values from -30 to infinite and average 11.2. *V4* indicates behavior characteristics of customer in the past two years. In cases when customer has not taken loan during this period, this variable will be in the *"Not Applied"* group; otherwise, seven other groups are created and the categories are: *V4_1, V4_2, ..., V4_8*.

V4_8 is decided to be a reference cell from the fact that this information is not available for these observations. Also, because the bad rate in this group is very close to the average one. *V4_4* also has bad rate close to the average and, because *V4_8* is a very small group, *V4_4* will also be the reference cell. The bad rate and distribution of *V4* can be analyzed below:

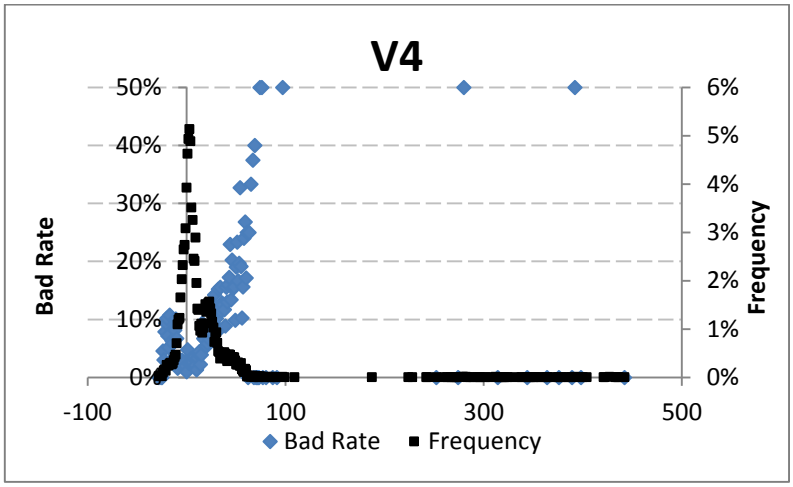

Figure 2.4 : **Bad rate and frequency of explanatory variable V4**

Variable V4						
Categories	**Band**	**# Bad**	**# Good**	**# Total**	**% Total**	**% Bad**
V4_1	-26 to -5	117	2592	2709	12%	4.3%
V4_2	-4 to 0	67	3345	3412	15%	2.0%
V4_3	1 to 10	243	8115	8358	36%	2.9%
V4_4	11 - 21	159	2938	3097	13%	5.1%
V4_5	22 - 27	156	1708	1864	8%	8.4%
V4_6	28 - 42	237	1678	1915	8%	12.4%
V4_7	43+	307	1059	1366	6%	22.5%
V4_8	Not Applied	24	421	445	2%	5.4%
Total		**1310**	**21856**	**23166**	**100%**	**6%**

Table 2.4 : **Characteristics summary of explanatory dummy variable V4**

Variable V5

It is a discrete variable starting from 0 to infinite, also is a behavior type. It mostly concentrates in lower values of *V5*, and the average is 22.2.

Six categories are created from *V5: V5_1, V5_2, ..., V5_6* and *V5_3* is the reference cell.

Variable V5

Categories	Band	# Bad	# Good	# Total	% Total	% Bad
V5_1	0 - 5	555	5120	5675	24%	9.8%
V5_2	6 - 10	233	3097	3330	14%	7.0%
V5_3	11 - 14	107	1973	2080	9%	5.1%
V5_4	15 - 29	222	5032	5254	23%	4.2%
V5_5	30 - 48	105	3088	3193	14%	3.3%
V5_6	49+	88	3546	3634	16%	2.4%
Total		**1310**	**21856**	**23166**	**100%**	**6%**

Table 2.5 : **Characteristics summary of explanatory dummy variable V2**

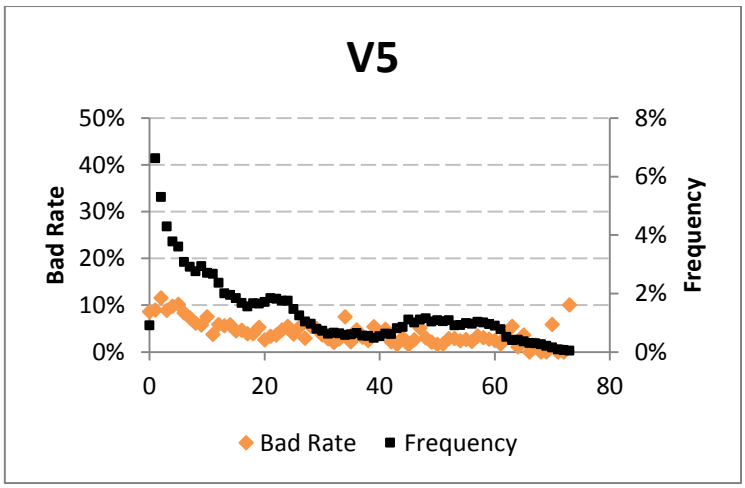

Figure 2.2 : **Bad rate and frequency of explanatory variable V5**

Variable V6

The last variable to be analyzed is *V6*. It is a categorical variable with 100 possible outcomes which were grouped in 6 different categories. *V6_4* is the neutral category. It is considered as demographic type. In the graph the categories were ordered by the bad rate.

Variable V6

Categories	# Bad	# Good	# Total	% Total	% Bad
V6_1	23	793	816	4%	2.8%
V6_2	157	3240	3397	15%	4.6%
V6_3	478	8590	9068	39%	5.3%
V6_4	209	3233	3442	15%	6.1%
V6_5	317	4631	4948	21%	6.4%
V6_6	126	1369	1495	6%	8.4%
Total	**1310**	**21856**	**23166**	**100%**	**6%**

Table 2.6 : **Characteristics summary of explanatory dummy variable V6**

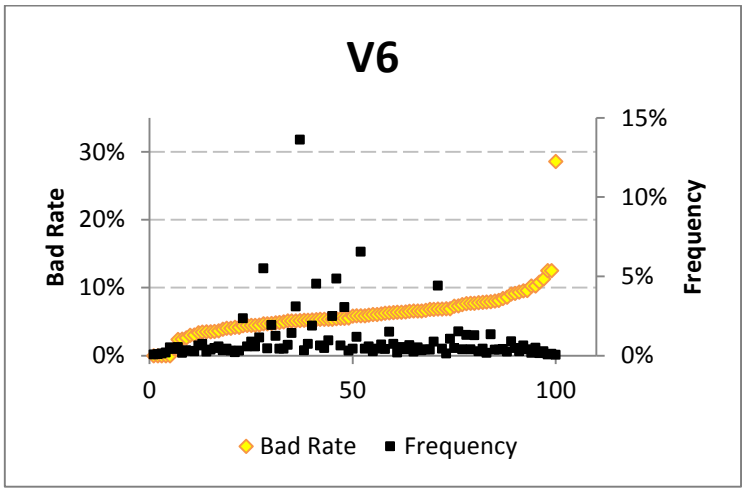

Figure 2.6 : **Bad rate and frequency of explanatory variable V6**

Chapter 3

Model comparison: Effect of excluding non-random part of the sample

3.1 – Logistic Regression: A Common Approach in the Financial Market

When credit scoring was first developed, statistical discrimination and classification were the only methods applied at that time and they remained the most important method by far. The method, offered by Fisher (1936), examines common classification dilemmas based on the discrimination methods, which could be viewed as a form of linear regression and more forms of regression models were continuously investigated by then. By far the most successful and common statistical method is logistic regression, which has less restrictive assumptions to guarantee their optimality and still lead to linear scoring rules (Lyn, Edelman, & Crook, 2002). One requirement for logistic regression is to have a large sample size, what is guaranteed in this study.

Logistic Regression versus Linear Regression

One can argue why not to use the linear regression instead. One practical reason is that it would generate predicted values more than 1 and less than 0, while the logistic regression outcomes can be used as probabilities since they are in the interval 0 and 1. It simplifies the use in the real world besides being more understandable for users. And theoretically, the use of linear regression combined with binary response would likely violate the assumptions of constant variance (heteroskedastic) and normal distribution (since the response is binary) of the error term.

The target

Logistic regression is built with a binary target. The target, in this case, reflects customers' classification according to the chance of default. In business, it is not interesting when customers pay back the loans very late since the company has extra expenses with collections, funding and it takes time to recover the debt. Therefore, the measure used to define the groups *"good"* and *"bad"* customers will be the number of days that the loan was paid back late and threshold is used to limit them.

A practice used by some credit risk score developers is to include an extra group of classification: *"neutral"* customers, as also mentioned by Hand & Henley, (1997, p.525). *"Neutral"* customers will also be defined by the number of days late in between the other two categories. The main idea is to have two extreme groups (*"good"* and *"bad"*) very well defined, giving the model more power to distinguish its difference when used for post prediction. In order to have this effect using logistic regression, the development sample is classified in those three categories and all *"neutral"* observations are disregarded of estimating the regression coefficients. However, this intuitive effect is not theoretically proved to be efficient and, moreover, may cause bias in the estimates. As a solution to the situation of a target that is not initially binary, the ordered logistic regression was chosen as the best option.

The ordered logistic outcome variable can be defined in different ways. Most of the references about the topic use the assumption that it should be derived from an unobserved and hypothetical continuous variable. However some authors, like Greene & Hensher (2009, p.83), Hosmer & Lemeshow (2000) and Dardanoni (2005, p.4) mention the possibility of using a latent dependent variable when applying ordinal logistic regression. Dardanoni state a theorem: *"If ε has a standard logistic distribution, the parameters $\boldsymbol{\beta}$ are the same in the latent regression and in the ordered logit models"*. In this paper the ordinal outcome arises by categorizing an observed discrete variable, number of days late, counting from the due to the payment date.

The second model will also use *"good"*, *"bad"* and *"neutral"* to classify customers but the development sample will be used in complete. It means that *'neutral'* will not be excluded from the estimation process.

The main idea is to fit *K-1* models, where *K* is the number of classification groups for the target. Each model will have a different constant but just one consistent estimator for the $\boldsymbol{\beta}$ coefficients. The formal comparison of this situation will be presented in next section.

3.2 – Theoretical Foundation

3.2.1 - Logistics regression

Logistic regression explains the relationship between a dependent binary variable and a set of explanatory variables. The estimated model can also 'predict' the outcome of a new observation for given values of the explanatory variables. Let Y denote the binary dependent variable being explained by J independent variables denoted by the vector $\boldsymbol{X}' = (X_1, X_2, \dots, X_J)$. As an example of interest in this thesis, suppose Y is a binary variable classifying the quality of the customers. Define $Y=1$ if the customer is a good customer as defined in previous chapter, and 0 otherwise. For a linear regression $\mathrm{E}(Y|\boldsymbol{X}) = \mu + \boldsymbol{X}'\boldsymbol{\beta}$ which is not restricted in range while the binary Y has a conditional expected value $\mathrm{E}(Y|\boldsymbol{X})$ between 0 and 1.

Logistic regression has been introduced to study the dichotomous data for its pleasant properties. Let $\pi(\boldsymbol{X})$ denote the probability that $Y=1$ given explanatory variable $\boldsymbol{X}$. It then follows $\mathrm{E}(Y|\boldsymbol{X}) = \pi(\boldsymbol{X})$. In the logistic regression model:

$$\pi(\boldsymbol{X}) = P(Y=1|\boldsymbol{X}) = \frac{\exp(\mu+\beta_1X_1+\beta_2X_2+\cdots+\beta_JX_J)}{1+\exp(\mu+\beta_1X_1+\beta_2X_2+\cdots+\beta_JX_J)} = \frac{\exp(\mu+\boldsymbol{X}'\boldsymbol{\beta})}{1+\exp(\mu+\boldsymbol{X}'\boldsymbol{\beta})} \quad (3.1)$$

$\pi(\boldsymbol{X})$ in equation 3.1 also satisfies the constraint which bond of probability Y given X should be between zero and 1.

One important study in logistic regression is the *logit transformation* where *"odds"* are introduced, i.e. $\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}$ means the probability of an event relative to the probability of the event not happening. The logit transformation is given in terms of $\pi(\boldsymbol{X})$:

$$g(\boldsymbol{X}) = ln\left(\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}\right) = \mu + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_JX_J = \mu + \boldsymbol{X}'\boldsymbol{\beta} \quad (3.2)$$

When **X** increases one unit, the odds ratio increases $\exp(\boldsymbol{\beta})$ unit.

Maximum likelihood Method

Maximum likelihood method is a commonly used method for fitting logistic regression models. *"Maximum likelihood yields values for the unknown parameters which maximize the probability of obtaining the observed set of data."* (Hosmer & Lemeshow, 2000). Likelihood function is constructed firstly as a function of the unknown parameters in an expression for the probability of the observed data, given the explanatory variables. That is:

$$l(\mu, \boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(X_i)^{y_i} [\, 1 - \pi(X_i) \,]^{1-y_i} \qquad (3.3)$$

The maximum likelihood estimator is defined as the value of the parameter which maximized the likelihood function (3.3). An easier option to maximize equation (3.3) is by using log:

$$L(\mu, \boldsymbol{\beta}) = ln[\, l(\hat{\mu}, \boldsymbol{\beta}) \,] = \sum_{i=1}^{n} \{\, y_i ln[\, \pi(X_i) \,] + (1 - y_i) ln[\, 1 - \pi(X_i) \,] \} \qquad (3.4)$$

Replacing $\pi(X_i) = \frac{e^{g(X_i)}}{1+e^{g(X_i)}}$,

$$L(\mu, \boldsymbol{\beta}) = \{\, y_i [\, g(X_i) \,] - ln[\, 1 + e^{\, g(X_i)} \,] \} \qquad (3.5)$$

Model Assessment

In the process of model assessment, the univariate Wald test is the first step to test the significance of the coefficients one at a time, in other words, whether the individual coefficient is zero or should remain in the model. The univariate Wald test statistics is defined as follow:

$$W_j = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)} \qquad (3.6)$$

These Wald statistics are approximately distributed as standard normal under the null hypothesis ($H_0 : \boldsymbol{\beta_j} = 0$). The significant value is considered as 0.05. The parameters will be selected according to the Wald statistics.

In the next step, comparison is made to test the fit of full model with J parameters and the reduced model with m parameters. The likelihood ratio test is applied here to differentiate

these two models under the null hypothesis: Ho: coefficients of the reduced variables are equal to 0.

$$G = -2ln\left[\frac{likelihood\ for\ reduced\ model\ with\ m\ parameters}{likelihood\ for\ model\ with\ J\ parameters}\right] \quad (3.7)$$

The statistics *G* is asymptotically chi-square distributed with $J - m$ degrees-of-freedom under the null hypothesis.

3.2.2 - Ordered logistic regression

Earlier sections presented the logistic regression model for binary dependent variables. This section introduces ordered logistic regression where the dependent variable is ordinal.

According to Hosmer & Lemeshow (2000), the three most widely used expressions for ordinal logistic regression are: the adjacent-category, the continuation-ratio and the proportional odds models. Considering that the software *Stata* is used in this paper to generate results and that it uses the proportional odds model for ordinal logistic regression, the focus of the explanation will be on this methodology.

Assume that the variable *Y* has *K* different outcomes, coded as *1,2,3,....,K* according to different categories of an unobserved continuous variable Y^*:

$$Y = \begin{cases} 1 & Y^* \leq \mu_1 \\ 2 & \mu_1 < Y^* \leq \mu_2 \\ & \dots \\ K & \mu_{K-1} < Y^* \end{cases} \quad (3.8)$$

where *Y* is the ordered response and the $\boldsymbol{\mu}$ is thresholds which defines the Y^*. The thresholds $\mu_1, \mu_2, \mu_3 \dots \mu_{k-1}$ should be greater than the previous threshold. μ_1 is normalized to 0 which gives one less parameter to be estimated. Consider a linear regression model for Y* :

$$Y^* = \boldsymbol{X}'\boldsymbol{\beta} + \varepsilon \quad (3.9)$$

where $\boldsymbol{X}$ is a vector of regressors, $\boldsymbol{\beta}$ is a vector of unknown regression parameters, and ε is a random unobserved disturbance term. The disturbance term is assumed independent of X. If ε is logistically distributed, then

$$P(Y \leq k|\boldsymbol{X}) = F(\mu_k - \boldsymbol{X}'\boldsymbol{\beta}) = \frac{exp(\mu_k - \boldsymbol{X}'\boldsymbol{\beta})}{1+exp(\mu_k - \boldsymbol{X}'\boldsymbol{\beta})}, k = 1,2,\dots,K-1, \quad (3.10)$$

$$g(\boldsymbol{X}) = ln\left[\frac{P(Y \leq k|\boldsymbol{X})}{P(Y > k|\boldsymbol{X})}\right] = ln\left[\frac{\Phi_1(\boldsymbol{X})+\Phi_2(\boldsymbol{X})+\cdots+\Phi_k(\boldsymbol{X})}{\Phi_{k+1}(\boldsymbol{X})+\Phi_{k+2}(\boldsymbol{X})+\cdots+\Phi_K(\boldsymbol{X})}\right] = \mu_k - \boldsymbol{X}'\boldsymbol{\beta} \quad (3.11)$$

Where *F()* is the cumulative distribution function of logistic distribution. Since y is discrete,

$$P(y_i = k|\boldsymbol{X}) = \Phi_k(\boldsymbol{X_i}) = F(\mu_k + \boldsymbol{X_i'}\boldsymbol{\beta}) - F(\mu_{k-1} + \boldsymbol{X_i'}\boldsymbol{\beta}) \quad (3.12)$$

For $k \leq K\text{-}1$ and

$$P(y_i = K) = \Phi_k(\boldsymbol{X_i}) = 1 - F(\mu_{K-1} + \boldsymbol{X_i'}\boldsymbol{\beta}) \quad (3.13)$$

Using these last expressions it is possible to define a likelihood for given sample of observations of y and X; whereby the unknown parameter vector $\boldsymbol{\beta}$ and the thresholds μ can be estimated with ML.

Maximum likelihood estimation for ordered logistic regression

Earlier maximum likelihood estimation was considered for the binary logistic regression model. The coefficients of the ordinal logistic model can also be estimated based on ML method. Here the likelihood is given by the expression

$$l(\mu, \boldsymbol{\beta}) = \prod_{i=1}^{n} [\, \Phi_1(\boldsymbol{X_i})^{Z_{1i}} \Phi_2(\boldsymbol{X_i})^{Z_{2i}} \dots \Phi_K(\boldsymbol{X_i})^{Z_{Ki}} \,] \quad (3.14)$$

where $(\boldsymbol{X_i}, y_i)$, $i = 1,2,\dots,n$ is a sample of n independent observations and vector $\boldsymbol{Z}' = (z_1, z_2, \dots, z_K)$ is created as K dimensional multinomial outcome where $z_k = 1$ if $y = k$ and $z_k = 0$ otherwise. The log-likelihood function is:

$$L(\mu, \boldsymbol{\beta}) = ln[\, l(\mu, \boldsymbol{\beta}) \,] = \sum_{i=1}^{n} \{Z_{1i} ln[\, \Phi_1(\boldsymbol{X_i}) \,] + Z_{2i} ln[\, \Phi_2(\boldsymbol{X_i}) \,] + \cdots + Z_{Ki} ln[\, \Phi_K(\boldsymbol{X_i}) \,]\}$$

(3.15)

The coefficients estimator $\hat{\mu}$ and $\widehat{\boldsymbol{\beta}}$ will be obtained by differentiating the last equation with respect to each of the coefficients and equalizing all the equations to zero.

Parallel regression assumption

The ordered logistic model is defined with different constants, however the same coefficient vector $\boldsymbol{\beta}$. According to Long (1997, p.141), this is the feature of the model under "parallel regression assumption" which should be examined in application of ordered model. In this paper, the Brant's (1990) test [Williams (2006, p.3)] is applied to test the parallel regression assumption in a straightforward expression:

$$\boldsymbol{\beta_1} = \boldsymbol{\beta_2} = \cdots = \boldsymbol{\beta_{K-1}} \qquad (3.16)$$

The null hypothesis equaling to the above equation is explained in new expression:

$$H_0 : \boldsymbol{\beta_k} - \boldsymbol{\beta_1} = 0,\ k = 2, \ldots, K\text{-}1 \qquad (3.17)$$

or by summarizing as:

$$H_0 : \boldsymbol{R\beta}^* = 0$$

where

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{-I} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{-I} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{-I} \end{bmatrix},\ \boldsymbol{\beta}^* = \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \boldsymbol{\beta_3} \\ \vdots \\ \boldsymbol{\beta_{K-1}} \end{bmatrix} \qquad (3.18)$$

The Wald statistic will be:

$$\chi^2[\,J \times (K-1)\,] = \left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}^*\right)' \left[\boldsymbol{R} \times \boldsymbol{Asy.Var}\left[\widehat{\boldsymbol{\beta}}^*\right] \times \boldsymbol{R}'\right]^{-1} \left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}^*\right) \qquad (3.19)$$

where the asymptotic covariance matrix contains blocks :

$$\boldsymbol{Asy.Var}\left[\widehat{\boldsymbol{\beta}}^{*}\right](k,l) = Est.Asy.Cov\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}},\widehat{\boldsymbol{\beta}}_{\boldsymbol{l}}\right]$$

$$= \left[\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \widehat{\Lambda}_{ik}\left(1-\widehat{\Lambda}_{ik}\right)\boldsymbol{X_i X_i'}\right]^{-\mathbf{1}} \times \left[\sum_{\boldsymbol{i=}}^{\boldsymbol{n}} \widehat{\Lambda}_{il}\left(1-\widehat{\Lambda}_{ik}\right)\boldsymbol{X_i X_i'}\right] \times \left[\sum_{\boldsymbol{i=}}^{\boldsymbol{n}} \widehat{\Lambda}_{il}\left(1-\widehat{\Lambda}_{il}\right)\boldsymbol{X_i X_i'}\right]^{-\mathbf{1}}$$

(3.20)

and $\widehat{\Lambda}_{ik} = \Lambda\left(\hat{\mu}_k + \boldsymbol{X_i'}\widehat{\boldsymbol{\beta}}_k\right)$. Under the null hypothesis (3.17) the Wald statistic (3.19) is approximately Chi-square distributed with *J*(K-1)* degrees of freedom.

Generalized ordinal logistic regression is introduced in as an alternative which in regards of the assumption violation in ordered logistic regression. Results and discussion about both parallel regression assumption and generalized logistic regression will be presented in next Chapter 4.

3.3 – Model comparison with data exclusion

As discussed previously, the use of logistic regression model excluding from the sample the loans defined as *"neutral"*, here named M1, should be investigated. The focus will be given to the probability of default, i.e. the conditional probability given loans are good or bad. The conditional probability for *"Y = bad"* is:

$$P\big(Y = Bad \big| Y \in (Good, Bad)\big)$$

$$= \frac{P(Y = Bad)}{P(Y = Good) + P(Y = Bad)}$$

$$= 1 \Big/ \left[1 + \frac{\exp(\mu_1 - \boldsymbol{X'\beta}) * (1 + \exp(\mu_2 - \boldsymbol{X'\beta}))}{1 + \exp(\mu_1 - \boldsymbol{X'\beta})}\right]$$

(3.21)

The probability is calculated given that *Y* can be either *"good"* or *"bad"*, since *"neutral"* was excluded from the sample.

The logistic regression model based on full sample is named as M2, which has another probability density function to calculate probability of *"bad"* applications:

$$P(Y = Bad) = \frac{1}{1+exp(\mu - \boldsymbol{X'\beta})} \qquad (3.22)$$

"Bad" is coded as Y=0 and *"good"* is coded as Y=1 in M1 and M2.

One more case, the probability of *"bad"* can also be calculated based on the ordered logistic regression model, M3. Under the consideration of including *"neutral"*, will the result still be equal to M1? In M3 the full sample is used. The probability density function in ordinal logistic regression is obtained as follows:

$$P(Y = Good) = P(\boldsymbol{\varepsilon} < \mu_1 - \boldsymbol{X'\beta}) = \frac{exp(\mu_1 - \boldsymbol{X'\beta})}{1+exp(\mu_1 - \boldsymbol{X'\beta})} \qquad (3.23)$$

$$P(Y = Neutral) = P(\mu_1 - \boldsymbol{X'\beta} \leq \boldsymbol{\varepsilon} < \mu_2 - \boldsymbol{X'\beta}) \qquad (3.24)$$

$$P(Y = Bad) = P(\boldsymbol{\varepsilon} \geq \mu_2 - \boldsymbol{X'\beta}) = 1 - \frac{exp(\mu_2 - \boldsymbol{X'\beta})}{1+exp(\mu_2 - \boldsymbol{X'\beta})} = \frac{1}{1+exp(\mu_2 - \boldsymbol{X'\beta})} \quad (3.25)$$

Y^* is considered as how many days late for the payment, so in ordered logistic regression *"good"* is coded as Y= 1. In order to make the problem more clear and understandable, a simple numeric example is demonstrated below. The column $P(Y = Bad|M1)$ gives the conditional probabilities according to model (3.23) to (3.25). The last column gives the unconditional probabilities. In the second last column, the value μ in the expression (3.22) is solved for such that it yields the conditional probability $P(Y = Bad|M1)$.

$\boldsymbol{X'\beta}$	μ_1	μ_2	$P(Y = Bad\|M1)$	μ in M2	$P(Y = Bad\|M3)$
0	3	5.5	0.0043	5.4555	0.0041
1	3	5.5	0.0123	5.3841	0.011
3	3	5.5	0.1317	4.8857	0.0759
4	3	5.5	0.4042	4.3882	0.1824
5	3	5.5	0.7600	3.8471	0.3775
6	3	5.5	0.9292	3.4255	0.6225
7	3	5.5	0.9785	3.1833	0.8176
8	3	5.5	0.9928	3.0722	0.9241
9	3	5.5	0.9975	3.0273	0.9707
10	3	5.5	0.9991	3.0101	0.989

Table 3.1: **Probability of *"bad"* in M1 and M3; Threshold in M2 derived from M1**

The purpose is to judge whether there exists a threshold in M2 that can generate the same probability of *"bad"* in M1. The $\boldsymbol{X'\beta}$, μ_1 and μ_2 are manually imputed, probability of *"bad"* customer in M1 is calculated according to Formula (3.21). Suppose the final probability is the same for M1 and M2, then there should exits a threshold μ in M2 which can keep the same solution as M1. The result of the threshold in the logistic regression M1 is computed in Table 3.1. As the example demonstrates, μ is not constant over the different values of $\boldsymbol{X'\beta}$, which indicates that the conditional probability of *"bad"* cannot be modelled using (3.25). As a result,

if the removal of *"neutral"* from the data set is not appropriately addressed, model estimates may give erroneous predictions of bad customers.

Furthermore, to keep *"neutral"* as target also causes the difference of prediction. The prediction result in M2 is presented in Table 3.1. Keeping *"Neutral"* as an outcome, produces lower probabilities of turning into *"bad"*.

An illustration with real data is followed to demonstrate the dimension of this difference.

3.4 – Case Illustration

As it was explained in previous section, the first model is obtained by fitting a binary logistic regression to the sample without the *"neutral"* part of applications and will be called M1. Therefore the target is *"good"* versus *"bad"*. Second model, M2, will keep the *"neutral"* in the sample and it will be applied logistic regression in its conventional way, with the binary target: *"bad"* versus *"not bad",* being *"not bad"* the result of *"good"* and *"neutral"* condensed into one target group. Finally, third model uses ordinal logistic regression, M3, keeping *"neutral"* group and fitting a 3-level ordinal target. First of all, M1, M2 and M3 will be built using the same variable set. The variable selection criterion is to select all variables that are statistically significant for at least one of the models. The result follows, expressed by the variables' coefficients and their standard errors in parenthesis.

Model Comparison			
	M1	M2	M3
V1_1	-1.355***	-1.216***	-1.066***
	(0.110)	(0.110)	(0.080)
V1_2	-0.462***	-0.424***	-0.344***
	(0.110)	(0.110)	(0.080)
V1_4	0.007	-0.056	0.283***
	(0.130)	(0.130)	(0.090)
V2_1	-1.315***	-1.228***	-1.008***
	(0.160)	(0.150)	(0.110)
V2_2	-0.774***	-0.785***	-0.576***
	(0.160)	(0.150)	(0.100)
V2_3	-0.474**	-0.457**	-0.347***
	(0.140)	(0.140)	(0.090)
V2_5	0.422*	0.354*	0.273**
	(0.170)	(0.170)	(0.100)
V2_6	0.599***	0.592***	0.269**
	(0.180)	(0.180)	(0.100)
V2_7	1.134***	1.124***	0.536***
	(0.210)	(0.200)	(0.100)
V3_2	-1.005***	-0.889***	-0.997***
	(0.170)	(0.170)	(0.100)
V4_1	0.675***	0.525**	0.882***
	(0.200)	(0.200)	(0.130)
V4_2	0.933***	0.805***	1.162***
	(0.220)	(0.210)	(0.130)
V4_3	0.235	0.162	0.440***
	(0.170)	(0.170)	(0.100)
V4_5	-0.502***	-0.348*	-0.537***
	(0.150)	(0.150)	(0.090)
V4_6	-0.977***	-0.815***	-0.911***
	(0.14)	(0.140)	(0.080)
V4_7	-1.994***	-1.751***	-1.535***
	(0.14)	(0.130)	(0.090)
V5_1	-0.360***	-0.299**	-0.266***
	(0.110)	(0.100)	(0.080)
Constant(cut1)	4.000***	4.060***	-3.900***
	(0.21)	(0.20)	(0.120)
Constant(cut2)			-2.599***
			(0.120)
N	14617	16180	16180
Chi2	1591.633***	1342.872***	2881.202***
BIC	5401.765	5843.164	14399.608
Pseudo R2	0.2360	0.1941	0.1694

* p<0.1, **p<0.05, ***p<0.01

Table 3.2: **Model Comparison in full explanatory dummy variables of M1 M2 and M3**

All models are fit using the software *Stata*. M3 has two constants, since there are three possible target outcomes. The number of observations is also different, since M1 is reduced and M2, M3 are full sample. M3 generates coefficients with lower standard error than M1 and M2. That is positive to M3 because the confidence intervals are more precise and so the estimates. Moreover, the difference between coefficients in M1 and M2 are smaller in general, when compared to M3.

Pseudo R^2 is higher in M1. This statistic can be compared between models because it follows the same criteria of calculation but it does not have the same effects as when it is used for linear regression (OLS regression). Pseudo R^2 here is calculated according to McFadden's methodology but does not mean how well the predictors explain the variance of the dependent variable, so it should be used with caution. The main focus is to correctly detect *"bad"* customers, so the probability of *"Y=bad"* will be analyzed.

One way of comparing the models is to compute the correct classification rate. To be able to do so, it uses the estimated probability of *"bad"*. This target category is defined with the same criteria for all models. All 23166 applications from the full sample were classified as (+) if the probability of *"not bad"* is more than 0.7 and as (-) otherwise. The criteria to set the cutoff point is based on personal experience and the business interest which considers that it is better to lose a good customer than to approve a bad one. All analyzes in this chapter from now uses the same full sample, since it better express the population. Results are displayed in table below.

M1 vs M2 vs M3 - Classification Rate

		M1	M2	M3
Sensitivity	P (+ | G)	96.55%	98.21%	98.09%
Specificity	P (- | B)	23.13%	15.73%	15.88%
Positive predictive value	P (G | +)	95.45%	95.11%	95.11%
Negative predictive value	P (B | -)	28.69%	34.45%	33.28%
Correctly Classified Rate		92.40%	93.54%	93.44%

Table 3.3: **Classification rate comparison of M1, M2 and M3**

Sensitivity is the probability of correctly classify *"not bad"* applications. All models have excellent result for sensitivity and highest comes from M2, but the difference is not big.

Specificity is the probability of correctly classify *"bad"* applications. M2 and M3 have much lower results than M1, however it is still too low. It means that, according to M1, around 23% of all "bad" applications are actually classified as *"bad".*

Positive predictive value indicates the probability of all applications classified as *"not bad"* happen to be really *"not bad".* Results are very similar for all models and quite good. M1 has the highest result for a small difference.

Negative predictive value is the opposite, indicates the probability of, among all applications classified as *"bad",* be actually *"bad".* M2 has the highest probability but it is still low. M1 has the lowest result what means that around 72% of all applications classified as *"bad"* are actually *"not bad".* Applying this result in real problems, it can cause the high denial rate of applications that would be successfully paid back.

To compare how different M1 classifies applications from M2 and M3, the cross classification tables are displayed below. The divergence for both cases is M1 stating as *"bad"* some applications that are *"not bad"* for the other two models. This difference is smaller between M1 and M3.

Cross Classification M1 vs M2

	M2		
M1	+	-	Total
+	22110	0	22110
-	458	598	1056
Total	22568	598	23166

Table 3.4: **Cross classification of M1 vs M2**

Cross Classification M1 vs M3			
	M3		
M1	+	-	Total
+	22110	0	22110
-	431	625	1056
Total	22541	625	23166

Table 3.5: **Cross classification of M1 vs M3**

M1 and M2 agree in 98.02% of the cases while M1 and M3 agree in 98.13%. This result is a reflex of the Classification Rate table, showing that M2 and M3 generate very low probabilities of *"bad",* so the result is less applications being classified in such way.

One assumption for M3 is the parallel regression assumption and, for this data, ordered regression model fails, as it is shown in the Brant test. It also indicates which of these variables are statistically considered to have different coefficient estimates. All variables were selected in the table below so that they can all be tested individually.

Brant Test of Parallel Regression Assumption

Variable	chi2	p>chi2	df
All	186.84	0.000	26
v1_1	3.22	0.073	1
v1_2	0.39	0.530	1
v1_4	13.28	0.000	1
v2_1	2.94	0.086	1
v2_2	3.53	0.060	1
v2_3	0.93	0.334	1
v2_5	0.43	0.511	1
v2_6	5.11	0.024	1
v2_7	10.19	0.001	1
v3_2	0.47	0.491	1
v4_1	5.52	0.019	1
v4_2	4.24	0.039	1
v4_3	4.35	0.037	1
v4_5	2.94	0.086	1
v4_6	1.04	0.308	1
v4_7	4.79	0.029	1
v5_1	3.67	0.055	1
v5_2	5.27	0.022	1
v5_4	4.02	0.045	1
v5_5	1.54	0.215	1
v5_6	2.11	0.146	1
v6_1	1.88	0.170	1
v6_2	0.32	0.574	1
v6_3	0.69	0.405	1
v6_5	3.29	0.070	1
v6_6	4.20	0.040	1

A significant test statistic provides evidence that the regression assumption has been violated.

Table 3.6: **Brant test of parallel regression assumption in full explanatory dummy variables of M3**

When the Brant test statistic is significant, in other words, that there is evidence that the regression assumption has been violated, it can be interpreted as we reject the hypothesis that the coefficients for different binary regressions are the same. The practical result of the test, considering 5% of significance level, is that M3 is not recommended for this data and it can be

the reason that this model does not identify well the *"bad"* group. Other techniques will be suggested as solutions for this problem in next chapter, followed by an empirical example comparing these options.

Chapter 4

New Model Motivation

In the previous chapter it was discussed the application of the very popular logistic regression model to build credit score. Also, it was demonstrated the formal comparison and an illustration to this model, the classic logistic regression and the ordered logistic model. However, are they the only appropriate techniques to be considered? The answer is no and some of them will be discussed in this chapter.

The target is based on a variable that express time. It is the number of days from the due date till the payment day. It can assume negative values, when the payment is done before the due date, or positive values expressing how late the customer paid back.

Based on the nature of the target, one good suggestion would be to use duration model, also called survival model.

Duration model estimates how long time an individual remains in a certain state or takes an action. It is commonly used in economics and biologic field and can be also applied in credit scoring.

However, duration model could not be used as it is since the dependent variable present censoring problems. Depending on the strategy established by the company, late payers turn to be charged by collection companies and the exact moment when the payment is done is missing or not correctly reported. This is called right censoring and it is common in duration models. Details about how to deal with this problem will not be discussed here, but further research can be found in (Aalen, 1978 and Nelson, 1972) about a suggested non-parametric technique to adjust censoring problems.

Another point to be discussed about the target is its behavior mechanism. The payment is mostly concentrated on due date. The reasons for payment before and after the due date are possibly different: payment in advance has no advantage and, on the other side, payment after the due date brings negative consequence (extra fees, difficulty to take future loans, etc.). This fact may require two different models, one for each of these periods with such different characteristics.

On the top of that, as mentioned in Chapter 1, reminder letters, sms and phone calls taken as a measure for motivating the payment probably cause unexpected patterns in the payment behavior. Besides that, it is likely that the collections strategy changes in the future and the model would not fit so well anymore due to different collections actions in different periods. To fit duration model taking all these details into consideration would end up in a complicated model to handle, requiring many adjustments, what is completely doable but it is more susceptible for model misspecification.

The focus of this study is to bring the most appropriate solution to be applied in real problems. Lack of labor resources, tight deadlines and multitask working environment are common characteristics existent in companies that should be considered in the model choice. Because all the reasons argued above, duration model will not be taken for further tests.

In Section 3.4 the illustration case failed in the parallel regression assumption. When this assumption is not met, there are some options:

- Collapse some levels of the dependent variable: *"neutral"* could be collapsed with *"good"* and use logistic regression model, as it was done in Chapter 3.

- Use the generalized ordered logistic model: is the direct suggestion by many authors when the parallel regression assumption fails. It estimates $K - 1$ constant terms, as in

ordered logistic regression but the difference is that, for each of the $K - 1$ combinations of the groups, it is estimated different coefficients for the independent variables.

- Use the partial proportional odds model: is very similar to the generalized ordered logistic model, but some coefficients can be the same and others can differ along the $K-1$ group combinations.

- Use multinomial logistic model: the structure is very similar to the generalized ordered logistic model: $K-1$ constant terms and different coefficients for each explanatory variable. The difference is the feature of the dependent variable, here, not ordered. Multinomial logistic model is an extension of the logistic regression and it is very flexible but much more complex and its interpretation is not as straightforward. In this study case it is preferable to choose other techniques suitable for ordered target that may be more appropriate and parsimonious (Williams, Multinomial Logit Models - Overview [PDF document], 2011)

Regarding the circumstances, the suggestion is to use the generalized ordered logistic model which is appropriate for the ordered target, solve the problem of the parallel regression assumption, has a simpler approach which is easier to be interpreted by business users and will probably be as accurate the last two techniques mentioned. When exists the possibility of transforming the target to binary, the logistic regression may be preferred.

In next chapter will be displayed an empirical comparison of the techniques just discussed here, that seems to be appropriate for the problem. Overview of these models are printed below.

4.1 - Generalized Ordered Logit Model

For some reasons, assumption violation exists in the ordinal model which influences the result of model assessment and explanation. Therefore, the Generalized Ordered Logit Model is introduced as an alternative to generate new coefficients for model fitting.

Under the rejection of null hypothesis from Brant's (1990) test, a suggested model derives from assumption. The generalized ordered logit model is given below:

$$g(\boldsymbol{X}) = ln\left[\frac{P(Y \leq k|\boldsymbol{X})}{1-P(Y \leq k|\boldsymbol{X})}\right] = ln\left[\frac{\Phi_1(\boldsymbol{X})+\Phi_2(\boldsymbol{X})+\cdots+\Phi_k(\boldsymbol{X})}{\Phi_{k+1}(\boldsymbol{X})+\Phi_{k+2}(\boldsymbol{X})+\cdots+\Phi_K(\boldsymbol{X})}\right] = \mu_k + \boldsymbol{X'\beta_k} \quad (4.1)$$

where

$$\Phi_k = P(Y = k) = P(Y \leq k|\boldsymbol{X}) - P(Y < k|\boldsymbol{X}) \quad (4.2)$$

$$P(y_i \leq k) = F(\mu_k + \boldsymbol{X'\beta_k}) = \frac{exp(\mu_k+\boldsymbol{X'\beta_k})}{1+exp(\mu_k+\boldsymbol{X'\beta_k})}, k = 1,2,\dots,K-1 \quad (4.3)$$

As the formula presents, the coefficients of the vector $\boldsymbol{\beta}$ in generalized ordered logistic regression are not constant as it is in ordered logistic regression, instead, the coefficients differ across all levels of *k*.

4.2 - Partial Proportional Odds Model

In partial proportional odds model, the coefficients $\boldsymbol{\beta}$ are not different for all levels *k*. It is a mixture of ordered logistic and generalized ordered logistic models: some coefficients contain the properties of ordered logit model which does not vary for all levels of *k*, while the others are different, as the generalized ordered logistic.

For example: coefficients $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are the same for all values of *k* while the coefficients for X_3 and X_4 are different:

$$P(y_i > k) = \frac{exp(\alpha_k + X'_{1i}\beta_1 + X'_{2i}\beta_2 + X'_{3i}\boldsymbol{\beta_{3k}} + X'_{4i}\boldsymbol{\beta_{4k}})}{1 + exp(\alpha_k + X'_{1i}\beta_1 + X'_{2i}\beta_2 + X'_{3i}\boldsymbol{\beta_{3k}} + X'_{4i}\boldsymbol{\beta_{4k}})} \quad (4.4)$$

Chapter 5

Application and Results

In this chapter, two models are developed using different methodology as in Chapter 3, with the aim to compare their performance, stability, efficiency and accuracy in prediction to distinguish customers according to the risk that they represent in not paying their loans back.

In order to test the stability of the model, the sample is randomly divided in two parts: *development*, with 70% of the observations and *validation* with the remaining 30%. The coefficients are estimated based on the development sample and the probability of default is post-estimated in the validation and development sample and results compared.

Variable selection

Thereafter, the model was developed starting with all available explanatory variables. The variable selection is not a well-defined process: there are different techniques to select variables for a model. All variables must be tested, combined in form of interactions, used as they are, or in form of dummies to express categories. Transformation in variables like applying logarithm or squared terms may avoid bias in the estimates caused by errors in functional form of the independent variables (Whitehead, 1999). However, many articles about Credit Scoring application support the use of categorized variables, as explained in Chapter 2, and this will be the choice for this paper.

The model with all variables included is called the *"full model".* Software will provide the model output with Wald test, which tests the significance of each predictor variable. It was chosen a p-value of less than 0.05 to indicate significance for the *"reduced model",* which will keep just significant variables. All possibilities should be tested by comparing the Wald statistics for coefficients and the *Chi*-square or *R*-squared statistic between the models.

However, when excluded, omitted variables can cause bias to the estimates (Whitehead, 1999) and the best way to deal with this problem is to perform the likelihood ratio test, which will check whether the full model brings improvement over the reduced model. If the improvement is not observed, the reduced model should be chosen since inclusion of irrelevant variables ends up in a poor model fit.

Wald and *LR* test can give different result. It is not clear in the statistical theory which of these tests is superior, but statisticians tend to prefer the *LR* test.

Multicolinearity

In logistic regression, there are no assumptions related to the distributions of the explanatory variables. However, problems with estimation can occur when the explanatory variables are highly correlated with one another. This is called multicolinearity (Whitehead, 1999). In practice, variables that one is expecting to be significant and is not then should be checked. The table below shows the correlation of the explanatory dummy variables that will be tested for all models. The general result meets the assumption of no highly correlated explanatory variables. The cases highlighted in red should be carefully analyzed during the process of variable selection but are still considered as normal.

	v1_1	v1_2	v1_4	v2_1	v2_2	v2_3	v2_5	v2_6	v2_7	v3_2	v4_1
v1_1	1										
v1_2	-0.090	1									
v1_4	-0.103	-0.128	1								
v2_1	0.202	0.099	-0.047	1							
v2_2	0.092	0.059	-0.027	-0.163	1						
v2_3	0.030	0.045	-0.030	-0.201	-0.159	1					
v2_5	-0.072	-0.030	0.003	-0.171	-0.136	-0.168	1				
v2_6	-0.101	-0.061	0.003	-0.185	-0.147	-0.181	-0.154	1			
v2_7	-0.121	-0.126	0.102	-0.205	-0.162	-0.200	-0.171	-0.184	1		
v3_2	-0.067	-0.041	-0.048	-0.192	-0.055	-0.007	0.059	0.105	0.083	1	
v4_1	0.098	0.030	-0.011	0.312	0.051	-0.029	-0.090	-0.104	-0.108	-0.258	1
v4_2	0.017	0.005	0.072	0.045	0.045	0.012	-0.027	-0.041	-0.029	-0.291	-0.151
v4_3	-0.027	-0.008	0.020	-0.098	-0.038	0.002	0.027	0.021	0.065	**-0.400**	-0.273
v4_5	-0.021	0.002	-0.029	-0.066	-0.001	0.011	0.019	0.044	0.004	**0.387**	-0.108
v4_6	-0.016	-0.018	-0.030	-0.073	-0.019	0.008	0.027	0.041	0.010	**0.393**	-0.109
v4_7	-0.021	-0.008	-0.006	-0.066	-0.011	0.014	0.039	0.012	-0.008	0.328	-0.091
v5_1	0.298	0.128	-0.054	**0.569**	0.247	0.010	-0.211	-0.232	-0.257	-0.242	0.279
v5_2	0.060	0.074	-0.010	-0.061	0.077	0.217	-0.009	-0.140	-0.185	-0.052	-0.027
v5_4	-0.140	0.009	-0.023	-0.190	-0.108	-0.043	0.174	0.187	-0.071	0.070	-0.122
v5_5	-0.108	-0.116	0.043	-0.159	-0.108	-0.102	0.037	0.165	0.204	0.119	-0.073
v5_6	-0.116	-0.144	0.057	-0.184	-0.135	-0.141	-0.031	0.064	**0.477**	0.144	-0.063
v6_1	-0.001	0.005	0.017	0.002	0.000	-0.019	0.006	-0.008	0.021	-0.006	-0.010
v6_2	0.006	-0.007	-0.009	-0.006	0.001	0.010	-0.023	-0.002	0.017	-0.010	-0.011
v6_3	-0.002	-0.004	-0.035	-0.010	-0.014	0.004	0.009	0.001	-0.007	-0.019	-0.004
v6_5	0.001	0.019	-0.024	0.011	0.006	0.000	0.010	0.000	-0.016	-0.020	-0.004
v6_6	0.013	0.013	0.002	0.019	0.017	-0.003	-0.004	-0.010	-0.008	0.015	-0.001

	v4_2	v4_3	v4_5	v4_6	v4_7	v5_1	v5_2	v5_4	v5_5	v5_6	v6_1
v4_2	1										
v4_3	-0.312	1									
v4_5	-0.123	-0.222	1								
v4_6	-0.125	-0.226	-0.089	1							
v4_7	-0.104	-0.188	-0.074	-0.075	1						
v5_1	0.110	-0.051	-0.073	-0.096	-0.087	1					
v5_2	0.051	0.017	-0.003	-0.007	-0.011	-0.233	1				
v5_4	-0.060	0.028	0.020	0.058	0.049	-0.309	-0.222	1			
v5_5	-0.044	-0.015	0.029	0.029	0.030	-0.228	-0.164	-0.217	1		
v5_6	-0.058	0.002	0.023	0.015	0.012	-0.246	-0.177	-0.234	-0.173	1	
v6_1	0.001	0.004	-0.018	0.002	-0.014	-0.003	-0.005	-0.010	0.006	0.018	1
v6_2	0.014	0.000	-0.010	0.005	-0.011	-0.014	0.006	0.002	0.006	0.003	-0.065
v6_3	-0.011	0.018	0.011	-0.015	-0.012	-0.013	0.009	0.029	-0.034	0.003	-0.088
v6_5	0.013	0.017	-0.006	-0.006	-0.005	0.013	-0.001	0.000	-0.004	-0.016	-0.073
v6_6	-0.011	0.001	-0.003	0.012	0.013	0.022	-0.004	-0.010	-0.008	0.003	-0.045

	v6_2	v6_3	v6_5	v6_6
v6_2	1			
v6_3	-0.191	1		
v6_5	-0.159	-0.216	1	
v6_6	-0.098	-0.134	-0.111	1

Table 5.1: **Correlation of explanatory dummy variables**

Model 1 – Generalized Ordered Logistic Model (GOL)

Quednau, Clogg and Shihadeh, Fahrmeir and Tutz, McCullagh and Nelder have proposed versions of the ordered choice models when there is no proportionality of odds across response categories. Fu and Williams provided a Stata program to estimate the Generalized Ordered Regression model (Greene & Hensher, 2009).

The Brant test was presented in details in Chapter 3 (table 3.6), so it is possible to check which variables were responsible for the failure of the parallel regression assumption.

Generalized ordered logistic regression is the first solution for this problem and estimates as many coefficients as the number of binary regressions, *K-1*. The final model is fit below:

Generalized Logistic Regression Model -- GOL

target3	Full	Reduced
1		
V1_1	-1.197***	-1.227***
	(0.110)	(0.110)
V1_2	-0.453***	-0.450***
	(0.110)	(0.110)
V1_4	-0.057	-0.06
	(0.120)	(0.120)
V2_1	-1.222***	-1.259***
	(0.160)	(0.160)
V2_2	-0.797***	-0.831***
	(0.150)	(0.150)
V2_3	-0.471***	-0.487***
	(0.140)	(0.140)
V2_5	0.392*	0.389*
	(0.170)	(0.170)
V2_6	0.706***	0.700***
	(0.190)	(0.180)
V2_7	1.309***	1.311***
	(0.230)	(0.230)
V3_2	-0.895***	-0.916***
	(0.160)	(0.160)
V4_1	0.581**	0.530**
	(0.190)	(0.190)
V4_2	0.864***	0.846***
	(0.200)	(0.200)
V4_3	0.207	0.19
	(0.150)	(0.150)
V4_5	-0.379**	-0.397**
	(0.140)	(0.140)
V4_6	-0.911***	-0.916***
	(0.140)	(0.140)
V4_7	-1.810***	-1.820***
	(0.130)	(0.130)
V5_1	-0.533***	-0.336**
	(0.160)	(0.110)
V5_2	-0.307*	
	(0.150)	
V5_4	-0.179	
	(0.150)	
V5_5	-0.335	-0.172
	(0.190)	(0.150)
V5_6	-0.434*	-0.265
	(0.210)	(0.170)
V6_1	0.442	
	(0.270)	
V6_2	0.117	
	(0.120)	
V6_3	0.066	
	(0.100)	
V6_5	-0.129	
	(0.100)	
V6_6	-0.247	
	(0.130)	
Constant	4.294***	4.155***
	(0.230)	(0.190)

target3	Full	Reduced
2		
V1_1	-1.016***	-1.021***
	(0.090)	(0.090)
V1_2	-0.364***	-0.363***
	(0.080)	(0.080)
V1_4	0.324***	0.321***
	(0.090)	(0.090)
V2_1	-0.960***	-0.972***
	(0.110)	(0.110)
V2_2	-0.537***	-0.551***
	(0.110)	(0.110)
V2_3	-0.333***	-0.345***
	(0.090)	(0.090)
V2_5	0.289**	0.302**
	(0.110)	(0.100)
V2_6	0.295**	0.319**
	(0.110)	(0.100)
V2_7	0.635***	0.655***
	(0.120)	(0.120)
V3_2	-0.970***	-0.974***
	(0.100)	(0.100)
V4_1	0.891***	0.882***
	(0.120)	(0.120)
V4_2	1.164***	1.153***
	(0.130)	(0.130)
V4_3	0.460***	0.447***
	(0.100)	(0.100)
V4_5	-0.581***	-0.590***
	(0.090)	(0.090)
V4_6	-0.955***	-0.958***
	(0.090)	(0.090)
V4_7	-1.506***	-1.507***
	(0.090)	(0.090)
V5_1	-0.221*	-0.275***
	(0.110)	(0.080)
V5_2	0.043	
	(0.100)	
V5_4	0.117	
	(0.100)	
V5_5	-0.105	-0.188**
	(0.110)	(0.080)
V5_6	-0.134	-0.219*
	(0.120)	(0.090)
V6_1	0.11	
	(0.150)	
V6_2	0.175*	
	(0.080)	
V6_3	0.011	
	(0.070)	
V6_5	0.012	
	(0.070)	
V6_6	-0.008	
	(0.100)	
Constant	2.506***	2.612***
	(0.140)	(0.120)

N	16180	16180
Chi2	3090***	3065***
BIC	14529.7	14419.7
Pseudo R2	0.1808	0.1793
Log Likelihood	-7003.1805	-7016.0228

* p<0.1, **p<0.05, ***p<0.01

Table 5.2: **Model Comparison for full and reduced explanatory dummy variables of GOL**

GOL - Classification Rate			
		Full	Reduced
Sensitivity	P(+|G)	97.95%	97.87%
Specificity	P(-|B)	17.63%	16.79%
Positive predictive value	P(G|+)	95.20%	95.15%
Negative predictive value	P(B|-)	34.07%	32.07%
Correctly Classified Rate		93.41%	93.28%

Table 5.3: **Classification rate comparison for full and reduced explanatory dummy variables of GOL**

Dummies that are significant for at least one binary regression are kept in the *"reduced"* model. Coefficients and standard errors for each variable are very close to each other in both versions of the model. From GOL classification rate table, the models present similar result in predicting *"bad"* and *"not bad"*. The log likelihood in model with reduced variables is -7016.0228. Applied the likelihood test, the result presents do not reject the null hypothesis which means the model with reduced variables contains the same information as the full model.

Model 2 – Partial Proportional Odds Model (PPO)

The Brant test identifies which variables are considered to have different coefficients for different binary regressions. In partial proportional odds regression, different coefficients are imposed just for variables that violate the parallel assumption. It works as an iterative process beginning with the GOL model and then a series of Wald tests to check if the coefficients are equal in the equations. The variable with least significance is constrained to have equal effects across equations. The model is re-estimated with the constraint and the process is repeated till there are no more variables that meet the parallel assumption (Williams, Generalized ordered logit / partial proportional odds models for ordinal dependent variables, 2006). Variables that do not meet the parallel assumption will not be constrained and have different coefficients estimates. In the end, a general Wald test is done and an insignificant test result indicates that the final model does not violate the parallel assumption.

The iterations for the *"full model"* are displayed below as an example. Note that the first step selected *V4_6* as the less significant variable for the Wald test (*p*-value = 0.7199) to be constrained. The following steps have lower p-values until the limit of 0.05, when variables are not constrained anymore and have different coefficients estimates.

Testing parallel lines assumption using the .05 level of significance...

Step 1: Constraints for parallel lines imposed for v4_6 (P Value = 0.7199)
Step 2: Constraints for parallel lines imposed for v6_2 (P Value = 0.5580)
Step 3: Constraints for parallel lines imposed for v2_5 (P Value = 0.4878)
Step 4: Constraints for parallel lines imposed for v3_2 (P Value = 0.4746)
Step 5: Constraints for parallel lines imposed for v6_3 (P Value = 0.4568)
Step 6: Constraints for parallel lines imposed for v1_2 (P Value = 0.3176)
Step 7: Constraints for parallel lines imposed for v5_5 (P Value = 0.2178)
Step 8: Constraints for parallel lines imposed for v5_6 (P Value = 0.3188)
Step 9: Constraints for parallel lines imposed for v6_1 (P Value = 0.1902)
Step 10: Constraints for parallel lines imposed for v5_4 (P Value = 0.1380)
Step 11: Constraints for parallel lines imposed for v2_3 (P Value = 0.1630)
Step 12: Constraints for parallel lines imposed for v4_5 (P Value = 0.0986)
Step 13: Constraints for parallel lines imposed for v1_1 (P Value = 0.0834)
Step 14: Constraints for parallel lines imposed for v2_1 (P Value = 0.0643)
Step 15: Constraints for parallel lines imposed for v2_2 (P Value = 0.2892)
Step 16: Constraints for parallel lines are not imposed for
v1_4 (P Value = 0.00043)
v2_6 (P Value = 0.00200)
v2_7 (P Value = 0.00006)
v4_1 (P Value = 0.00000)
v4_2 (P Value = 0.00027)
v4_3 (P Value = 0.00000)
v4_7 (P Value = 0.00002)
v5_1 (P Value = 0.00001)
v5_2 (P Value = 0.00363)
v6_5 (P Value = 0.03176)
v6_6 (P Value = 0.01374)

Table 5.4: **Parallel regression assumption test in PPO**

The final test is automatically generated and shows that the final model meets the parallel assumption.

Wald test of parallel lines assumption for the final model:

(1) [1]v4_6 - [2]v4_6 = 0
(2) [1]v6_2 - [2]v6_2 = 0
(3) [1]v2_5 - [2]v2_5 = 0
(4) [1]v3_2 - [2]v3_2 = 0
(5) [1]v6_3 - [2]v6_3 = 0
(6) [1]v1_2 - [2]v1_2 = 0
(7) [1]v5_5 - [2]v5_5 = 0
(8) [1]v5_6 - [2]v5_6 = 0
(9) [1]v6_1 - [2]v6_1 = 0
(10) [1]v5_4 - [2]v5_4 = 0
(11) [1]v2_3 - [2]v2_3 = 0
(12) [1]v4_5 - [2]v4_5 = 0
(13) [1]v1_1 - [2]v1_1 = 0
(14) [1]v2_1 - [2]v2_1 = 0
(15) [1]v2_2 - [2]v2_2 = 0

chi2(15) =   21.82
Prob > chi2 =   0.1125

An insignificant test statistic indicates that the final model
does not violate the proportional odds/ parallel lines assumption

Table 5.5: **Wald test of parallel regression assumption in PPO**

And the final partial proportional odds model output is displayed below:

Partial Proportional Odds Model -- PPO

target3	Full	Reduced
1		
V1_1	-1.053***	-1.232***
	(0.090)	(0.100)
V1_2	-0.374***	-0.376***
	(0.080)	(0.080)
V1_4	-0.024	-0.034
	(0.120)	(0.120)
V2_1	-0.998***	-1.252***
	(0.110)	(0.130)
V2_2	-0.576***	-0.806***
	(0.100)	(0.130)
V2_3	-0.349***	-0.366***
	(0.090)	(0.090)
V2_5	0.294**	0.306**
	(0.100)	(0.100)
V2_6	0.716***	0.753***
	(0.160)	(0.160)
V2_7	1.306***	1.349***
	(0.200)	(0.200)
V3_2	-0.970***	-0.973***
	(0.100)	(0.100)
V4_1	0.396*	0.413**
	(0.160)	(0.160)
V4_2	0.712***	0.720***
	(0.180)	(0.180)
V4_3	0.068	0.078
	(0.120)	(0.120)
V4_5	-0.558***	-0.567***
	(0.090)	(0.090)
V4_6	-0.954***	-0.956***
	(0.080)	(0.080)
V4_7	-1.866***	-1.868***
	(0.110)	(0.110)
V5_1	-0.557***	-0.287***
	(0.120)	(0.080)
V5_2	-0.236	
	(0.130)	
V5_4	0.082	
	(0.100)	
V5_5	-0.132	-0.187*
	(0.110)	(0.080)
V5_6	-0.166	-0.223*
	(0.120)	(0.090)
V6_1	0.141	
	(0.150)	
V6_2	0.168*	
	(0.080)	
V6_3	0.016	
	(0.060)	
V6_5	-0.158	
	(0.100)	
V6_6	-0.264*	
	(0.130)	
Constant	4.200***	4.207***
	(0.150)	(0.130)

target3	Full	Reduced
2		
V1_1	-1.053***	-1.019***
	(0.090)	(0.090)
V1_2	-0.374***	-0.376***
	(0.080)	(0.080)
V1_4	0.322***	0.320***
	(0.090)	(0.090)
V2_1	-0.998***	-0.971***
	(0.110)	(0.110)
V2_2	-0.576***	-0.552***
	(0.100)	(0.100)
V2_3	-0.349***	-0.366***
	(0.090)	(0.090)
V2_5	0.294**	0.306**
	(0.100)	(0.100)
V2_6	0.291**	0.310**
	(0.110)	(0.100)
V2_7	0.632***	0.647***
	(0.120)	(0.120)
V3_2	-0.970***	-0.973***
	(0.100)	(0.100)
V4_1	0.904***	0.893***
	(0.120)	(0.120)
V4_2	1.171***	1.165***
	(0.130)	(0.130)
V4_3	0.465***	0.457***
	(0.100)	(0.090)
V4_5	-0.558***	-0.567***
	(0.090)	(0.090)
V4_6	-0.954***	-0.956***
	(0.080)	(0.080)
V4_7	-1.498***	-1.505***
	(0.090)	(0.090)
V5_1	-0.210*	-0.287***
	(0.110)	(0.080)
V5_2	0.033	
	(0.100)	
V5_4	0.082	
	(0.100)	
V5_5	-0.132	-0.187*
	(0.110)	(0.080)
V5_6	-0.166	-0.223*
	(0.120)	(0.090)
V6_1	0.141	
	(0.150)	
V6_2	0.168*	
	(0.080)	
V6_3	0.016	
	(0.060)	
V6_5	0.014	
	(0.070)	
V6_6	-0.005	
	(0.100)	
Constant	2.533***	2.614***
	(0.140)	(0.120)

N	16180	16180
Chi2	3068.472***	3055.868***
BIC	14406.168	14341.24
Pseudo R2	0.1795	0.1787
Log Likelihood	-7014.0992	-7020.4014

* p<0.1, **p<0.05, ***p<0.01

Table 5.6: **Model Comparison for full and reduced explanatory dummy variables of PPO**

PPO - Classification Rate			
		Full	Reduced
Sensitivity	P(+|G)	98.01%	97.97%
Specificity	P(-|B)	17.10%	16.26%
Positive predictive value	P(G|+)	95.17%	95.13%
Negative predictive value	P(B|-)	33.94%	32.47%
Correctly Classified Rate		93.43%	93.35%

Table 5.7: **Classification rate comparison for full and reduced explanatory dummy variables of PPO**

PPO derives out similar solutions as GOL. Different variables selection leads to similar results. By reducing variables, no loss in information happens when comparing *"full"* and *"reduced model"*. Note that constrained dummies in the iteration process have the same coefficient estimates.

Comparing results

Generalized ordered logistic model (GOL), partial proportional odds model (PPO) and the model presented in Chapter 3, logistic regression of the sample with *"neutral"* applications excluded (M1) and logistic regression model in its classic way (M2) will be compared empirically.

Classification Statistics - Development Sample					
		M1	M2	GOL	PPO
Sensitivity	P(+|G)	96.51	99.67	97.86	97.97
Specificity	P(-|B)	**24.48**	4.94	17.45	16.79
Positive predictive value	P(G|+)	95.54	94.62	95.21	95.18
Negative predictive value	P(B|-)	29.50	46.88	32.78	33.05
Correctly Classified Rate		92.45	94.33	93.34	93.40

Table 5.8: **Classification statistics comparison of M1, M2, GOL and PPO in development sample**

Classification Statistics - Validation Sample					
		M1	M2	GOL	PPO
Sensitivity	P(+|G)	96.66	99.61	97.87	97.98
Specificity	P(-|B)	**20.05**	4.76	15.29	15.04
Positive predictive value	P(G|+)	95.23	94.53	95.02	95.01
Negative predictive value	P(B|-)	26.67	42.22	30.35	31.09
Correctly Classified Rate		92.28	94.19	93.16	93.24

Table 5.9: **Classification statistics comparison of M1, M2, GOL and PPO in validation sample**

The classification table above shows that the results are very similar to all models. The only relevant difference is in the higher specificity and lower negative predictive value of M1 what indicates that M1 tend to produce higher probabilities of *"bad"*. One reason can be that the rate of "bad" applications in the development sample is higher compared to the other models (since it is developed with just *"bad"* and *"good"* applications). However, to produce the classification table all applications were pre-scored so the comparison is valid.

The correlation matrix of the predicted probability of *"bad"* indicates high correlation between the models prediction.

Correlation Matrix for Predicted Y="bad"				
	M1	M2	GOL	PPO
M1	1.0000			
M2	0.9889	1.0000		
GOL	0.9870	0.9913	1.0000	
PPO	0.9872	0.9903	0.9963	1.0000

Table 5.10: **Correlation of prediction *"bad"* in M1, M2, GOL and PPO**

Another way to assess the performance, commonly used for credit scoring models is the KS - Kolmogorov-Smirnov statistic (Andrade & Oliveira, 2012). It measures the model ability to distinguish *"bad"* and *"not bad"* groups. The highest, the best. It is calculated by maximizing the difference between the cumulative distributions of *"bad"* and *"not bad"* groups. KS higher than 50% is not common for credit scoring for new customers, but it is for behavior models.

Twenty groups were created for all models results in a way that each has 5% of the sample applications. The table below has the KS results for all models, and M1 has the best result.

	KS
M1	0.5227
M2	0.5189
GOL	0.5189
PPO	0.5180

Table 5.11: **Kolmogorov-Smirnov statistic of M1, M2, GOL and PPO**

A way to represent KS graphically is plotting the cumulative distributions.

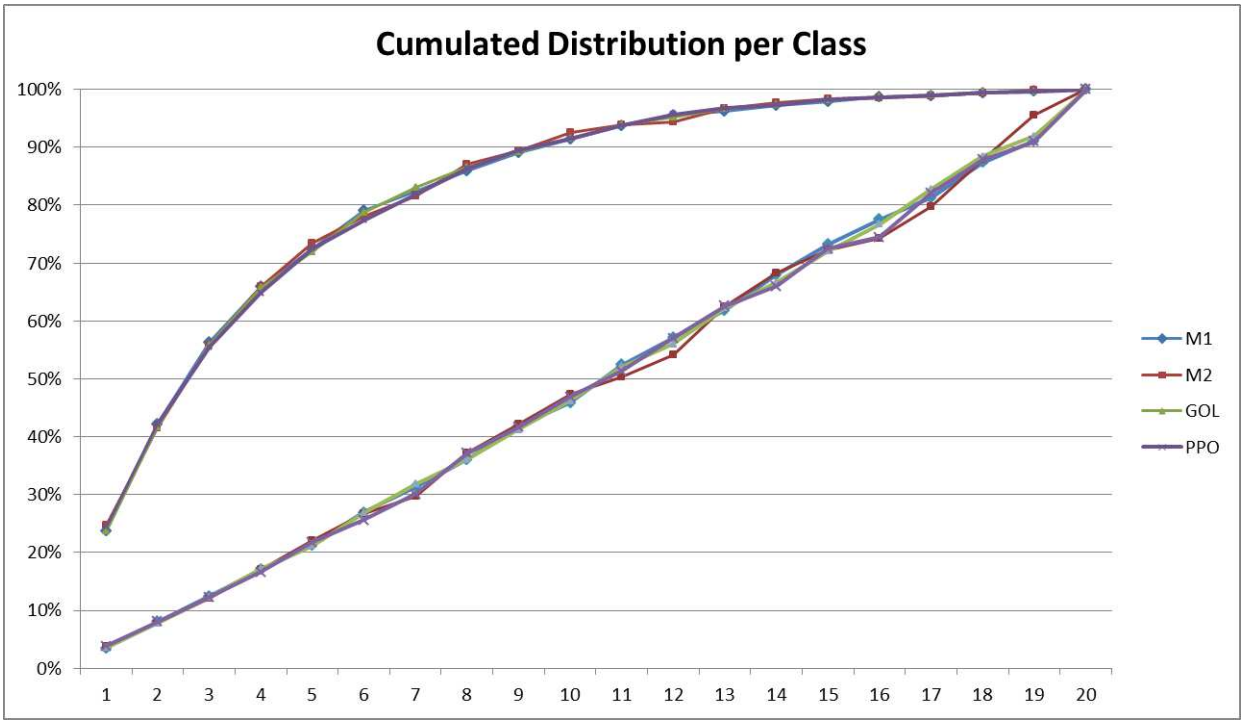

Figure 5.1: **Cumulated distribution per class of M1, M2, GOL and PPO**

The top lines are the *"bad"* cumulated distributions and the bottom is the *"not bad"*. All models have very similar distribution. KS is the maximum distance between those lines and it happens around 6th and 7th classes. This graph also expresses the classification table result, showing that many *"not bad"* customers are classified as *"bad"*. The ideal would be if the *"not bad"* cumulated distribution was curved to the opposite direction as the *"bad"*.

Mahalanobis distance is also used to evaluate model performance in credit score (Andrade & Oliveira, 2012). It measure how *"bad"* and *"not bad"* customers are distant by their score mean. The result is coherent with KS: M1 distinguish the groups better.

	MD
M1	0.8117
M2	0.7895
GOL	0.7866
PPO	0.7863

Table: 5.12: **Mahalanobis distance in M1, M2, GOL and PPO**

The conclusion taken from this paper and for the data used is that M1 can distinguish better the two groups *"bad"* and *"not bad"* when compared to M2, GOL and PPO, which have very similar results. However it also has the highest rate of misclassification for *"not bad"* customers. Thus, for a conservative type of business, M1 is the most appropriate among the four options.

From the theoretical point of view, M1 is not appropriate, since it excludes a non-random part of the sample. So the best shot is M2 for having slightly better results than the other two options and also for being a very easy solution.

However, if the purpose is to keep more than two target outcomes, GOL has slightly better results than PPO but both could be equally used. In this case it is a matter of preference of the

analyst. One situation that these models could be applied is, for example, for market purposes to send a campaign for *"neutral"* customers.

The ideal is to fit different models and compare them before implementation because models can behave in a different way when applied to different data.

Chapter 6

Conclusions

Credit scoring models can be developed using different techniques. Predictor variables are preferred to be categorized when applied for Credit Scoring, including continuous variables, as explained in Chapter 2.

In this paper it is developed Credit Scoring models using different techniques. The first approach is a very common use of the Logistic Regression when exist three possible outcomes: the *"neutral"* outcome is deleted from the development sample and model is fit, M1. Problems can arise from it, and it was shown that the structure of the probability function is different from the classic logistic regression, M2, and the ordered logistic regression, M3. In the case illustration, M1 shows to be more conservative: higher rate of applications being classified as *"bad"*. Besides that the three models had very similar results.

Ordered logistic model violated the parallel assumption so options were discussed for the case when more than two outcomes are needed. Generalized ordered logistic, GOL, and partial proportional odds, PPO, models are appropriate solutions and were compared in an empirical example. GOL has an easiest approach so is preferable.

If the company mainly focuses on prevention of default, M1 model will be the best opt for conservative business. It differs more in *"bad"* customers; meanwhile, it maintains a simply and understandable operability. A low percentage in default is always the crucial topic for a company. However, if the company would like to investigate potential *"good"* customers then GOL is recommended for having multinomial target. The behavior of *"neutral"* customers can be analyzed and compared with *"good"* and *"bad"*. Relevant campaigns can be raised to aim on intermediate customers which promote *"neutral"* into *"good"* and profits more from granting

higher loan or identifies the middle part clearly as *"bad"* and prevents the loss from default. The purpose and policy of company decides the focus of the business and also influences the choice of model. The information of how well the characteristics behave is basis for developing various promotions in business.

Statistical models, as presented in this paper, can show the contribution of each variable for the prediction though the coefficients estimators what is clearer to interpret. However, for further discussion, there are other non-statistical options to build a credit scoring like decision trees, neural networks, expert systems, genetic algorithms among others that can be used. These alternatives should be carefully analyzed because may easily turn to be unstable: fit very well in the development sample but out of sample, when applied in the real business, show different performance.

References

Aalen, O. O. (1978). Non Parametric Inference for a Family of Counting Processes. *The Annals of Statistics, 6*(4), 701–726.

Andrade, F. W., & Oliveira, J. G. (2012). Comparação entre Medidas de Performance de Modelos de Credit Scoring. *Revista Tecnologia de Crédito*.

Arellano, M. (2008). Duration Models [PDF document]. Retrieved 05 09, 2012, from http://www.cemfi.es/~arellano/duration-models.pdf

Bank of Mauritius. (2003, December). *Guideline on Credit Risk Management.* Retrieved April 14, 2012, from Bank of Mauritius: https://www.bom.mu/pdf/Legislation_Guidelines_Compliance/Guidelines/Guideline_on_Credit_Risk_Management.pdf

Bofondi, M., & Lotti, F. (2006, April 13). *Innovation in the Retail Banking Industry: the Diffusion of Credit Scoring*. Retrieved 05 09, 2012, from CommuniGate Pro: https://mail.sssup.it/~lotti/Bofondi_Lotti.pdf

Bolton, C. (2009). *Logistic Regression and its application in credit scoring*. University of Pretoria. Pretoria. Retrieved April 18, 2012 from http://upetd.up.ac.za/thesis/available/etd-08172010-202405/unrestricted/dissertation.pdf: (Unpublished master's thesis).

Burns, R., & Burns, R. (2008). *Chapter 24: Logistical Regression.* Retrieved April 17, 2012, from Business Research Methods and Statistics Using SPSS. Advance Online Publication.:

http://www.uk.sagepub.com/burns/website%20material/Chapter%2024%20-%20Logistic%20regression.pdf

Dardanoni, V. (2005). *Multivariate Ordered Logit Regressions.* University of Palermo, Department of Economics, Palermo. Retrieved April 19, 2012 from http://www.cemmap.ac.uk/forms/dardanoni_paper.pdf.

Greene, W. H., & Hensher, D. A. (2009). *Modeling Ordered Choices.* Cambridge: Cambridge University Press.

Hand, D. J., & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review. *J. R. Statist. Soc., 160*(3), 523-541.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (Second ed.). New York: John Wiley & Sons.

Kočenda, E., & Vojtek, M. (2009, Dec). *Default Predictors and Credit Scoring Models for Retail Banking.* Retrieved 05 2012, from CESIFO: http://www.ifo.de/pls/guestci/download/CESifo%20Working%20Papers%202009/CESifo%20Working%20Papers%20December%202009/cesifo1_wp2862.pdf

Long, J. S., & Freese, J. (2001). *Regression Models for Categorical Dependent Variables Using STATA.* College Station, TX: Stata Press.

Lyn, T. C., Edelman, D. B., & Crook, J. N. (2002). *Credit Scoring & Its Applications.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Nelson, W. (1972). Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics, 14*(4), 945–965.

Norušis, M. (2011). *IBM SPSS Statistics 19 Advanced Statistical Procedures Companion.* Upper Saddle River, NJ: Prentice Hall.

Pohlmann, J. T., & Leitner, D. W. (2003). A Comparison of Ordinary Least Squares and Logistic Regression. *Ohio Journal of Science*, 118-125.

Sullivan, A., & Sheffrin, S. M. (2003). *Economics: Principles in action.* Upper Saddle River, NJ: Pearson Prentice Hall.

Whitehead, J. (1999). *Introduction to Logistic Regression.* Retrieved April 14, 2012, from Appalachian State University: http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&ved=0CEEQFjAC&url=http%3A%2F%2Fwww.appstate.edu%2F~whiteheadjc%2Fservice%2Flogit%2Flogit.ppt&ei=HL6FT-vxBvL24QSGw8D8Bw&usg=AFQjCNELGe2isSPlyN-Id3FV2sPg7otXYg

Williams, R. (2006). Generalized ordered logit / partial proportional odds models for ordinal dependent variables. *The Stata Journal*, 58 - 82.

Williams, R. (2011). Multinomial Logit Models - Overview [PDF document]. Notre Dame, IN. Retrieved 05 09, 2012, from http://www.nd.edu/~rwilliam/stats2/l92.pdf

Appendix

M1 vs M2 vs M3 - Classification Table

	M1		M2		M3		
	Classified		Classified		Classified		
TRUE	+	-	+	-	+	-	Total
G	21103	753	21464	392	21439	417	21856
B	1007	303	1104	206	1102	208	1310
Total	22110	1056	22568	598	22541	625	23166

M1 - Development

	TRUE		
Classified	G	B	Total
+	14736	688	15424
-	533	223	756
Total	15269	911	16180

M1 - Validation

	TRUE		
Classified	G	B	Total
+	6367	319	6686
-	220	80	300
Total	6587	399	6986

M2 - Development

	TRUE		
Classified	G	B	Total
+	14992	763	15755
-	277	148	425
Total	15269	911	16180

M2 - Validation

	TRUE		
Classified	G	B	Total
+	6472	341	6813
-	115	58	173
Total	6587	399	6986

GOL - Development

	TRUE		
Classified	G	B	Total
+	14943	752	15695
-	326	159	485
Total	15269	911	16180

GOL - Validation

	TRUE		
Classified	G	B	Total
+	6447	338	6785
-	140	61	201
Total	6587	399	6986

PPO - Development

	TRUE		
Classified	G	B	Total
+	14959	758	15717
-	310	153	463
Total	15269	911	16180

PPO - Validation

	TRUE		
Classified	G	B	Total
+	6454	339	6793
-	133	60	193
Total	6587	399	6986