Default Predictors in Credit Scoring - Evidence from France's Retail Banking Institution

Ha Thu Nguyen

EconomiX
http://economix.fr

Document de Travail
Working Paper
2014-26

Default Predictors in Credit Scoring - Evidence from France's Retail Banking Institution

Ha-Thu Nguyen

UMR 7235

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

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

Default Predictors in Credit Scoring - Evidence from France's Retail Banking Institution

Ha-Thu Nguyen[1]

May 2014

Abstract.

The aim of this paper is to present the set-up of a behavioral credit-scoring model and to estimate such a model using the auto loan data set of one of the largest multinational financial institutions based in France. We rely on a logistic regression approach, which is commonly used in credit scoring, to construct a behavioral scorecard. A detailed description of the model building process is provided, as are discussions about specific modeling issues. The paper then uses a number of quantitative criteria to identify the model best suited to modeling. Finally, it is demonstrated that such model possesses the desirable characteristics of a scorecard.

Keywords. Auto Loans, Credit Risk, Credit Scoring, Logistic Regression.

JEL Classification. G3, C51, C52.

1. Introduction

"Credit-scoring technologies have served as the foundation for the development of our national markets for consumer and mortgage credit, allowing lenders to build highly diversified loan portfolios that substantially mitigate credit risk. Their use also has expanded well beyond their original purpose of assessing credit risk. Today they are used for assessing the risk-adjusted profitability of account relationships, for establishing the initial and ongoing credit limits available to borrowers, and for assisting in a range of activities in loan servicing, including fraud detection, delinquency intervention, and loss mitigation. These diverse applications have played a major role in promoting the efficiency and expanding the scope of our credit-delivery systems and allowing lenders to broaden the populations they are willing and able to serve profitably."

Alan Greenspan, U.S. Federal Reserve Chairman, October 2002

Speech to the American Bankers Association[2]

Given the key role played by credit in all economies, evaluating its associated risk is a crucial issue. The traditional way to evaluate credit risk, which relies on human analysis of a borrower's financials and future prospects, has been subject to critics. Various credit-related biases indeed exist, making human analysis less reliable. [3] In particular, agents tend to be too confident about their knowledge (**Alpert and Raiffa** (1982)). In addition, as shown by **Barber and Odean** (1999), agents' overconfidence bias is aggravated by the importance of the task, and by the fact that they tend to neglect past failures and focus on past successes. **Libby** (1975) shows that judgmental decision-making based on a batch of ratios allows bankers to predict firms' bankruptcy in a three-year horizon with a 74% accuracy, whereas using a simple method based solely on the liability/asset ratio actually

[1] EconomiX - CNRS – University of Paris West Nanterre La Défense, France
Email : hathunguyen12@gmail.com

I am heartily thankful to my supervisor, Valérie Mignon, and some anonymous reviewers, for valuable comments on earlier drafts of this paper. The providers of the data also requested anonymity. Therefore, I would like to thank them so such.

[2] http://www.federalreserve.gov/BoardDocs/Speeches/2002/20021007/default.htm

[3] See **Falkenstein et al.** (2000) among others.

improves this percentage. This result is in line with **Meehl** (1954), according to whom the human capacity to integrate complex and diverse sources of information is questionable. Anecdotal feedback (as opposed to statistical feedback) is given in lending institutions, making improving inductive reasoning difficult (**Nisbet, Krantz, Jepson, and Fong** (1982)).

The method of credit scoring is an attempt to address these concerns. Credit scoring is the methodology that scores accounts as to their likelihood of the outcome of, for example, becoming delinquent. It is often performed by lenders and financial institutions to access a person's credit-worthiness. Consequently, credit scoring relies much more on a quantitative model in which the criteria as well as their relative importance are well defined. This is of course not to say that credit scoring does not need human judgment, but it leaves less room for human error. For instance, in some retail finance, the sole results of a scoring model cannot be used for decision-making purpose. Compliance departments always require at least some degree of human judgment in the decision making process.

As a matter of fact, credit scoring was first used in the 1960s and is still widely used by banks and lending institutions for several reasons. First, credit scores help automating manual decisions and are based on the following statistical principle: If past trends indicate towards future trends, then a score leads to better decisions than manual ones because there is no subjectivity involved for the majority of decisions. Second, scores are also cost-effective because they reduce turnaround time, and potentially, headcount requirement. As a result, credit scores help lenders in assessing risk more fairly because they are consistent and objective. Third, consumers also benefit from this method: No matter who the clients are, their credit score only reflects their probability to repay debt obligation, based on their past credit history and current credit status. Thus, credit scores improve performance substantially while keeping costs down. Progress in technology and in the sharing of information, namely the establishment of Credit Bureau in the United States and similar institutions in other countries, have contributed to the rise of credit scoring in every kind of lending (**Altman and Saunders**, 1998).

The use of credit scores presents undoubtedly the greatest value in the decision-making process, when the underwriter must define clear scenarios based on scores and policies. The score allows the lender to answer questions such as whether to accept or reject a loan, or what are the maximum loan amounts, the applicable interest rates, the terms of the loan, etc. Credit scores may have different types, depending on where and how they are used. Some of these types are application scores, behavior scores, collections scores, fraud scores, etc. These names differ by the emphasis placed on different aspects of the scores, such as the source of information, the tasks being performed or what is being measured.

Credit scoring is an applied domain and has seen regular contribution from various authors. Early treatments of the scoring problems are **Bierman and Hausman** (1970) and **Dirickx and Wakeman** (1976), as well as **Srinivasan and Kim** (1987) among others. Most credit scoring models use past data to estimate risk for new applicants by assuming that past risk correctly predicts future risk. The number of factors taken into account has significantly increased with later models. **Hand** (1997) and **Thomas, Edelman and Crook** (2002) present credit scoring and its application in various lending activities. **Hand** (2001) also handles the issue of evaluating the effectiveness of credit scores ("scoring the score").

Various statistical methodologies have been investigated to construct credit scoring models over the last 50 years. With time, academic researchers focused more and more on credit scoring techniques,

then replaced the first oversimplified univariate analysis. **Beaver** (1967) and **Altman** (1968)[4] examined a multiple discriminant analysis (MDA) by using financial ratios as predictors of failure.

Several years later, the MDA was still the dominant statistical technique, and was further investigated by **Deakin** (1972), **Edmister** (1972), **Blum** (1974), **Altman et al.** (1977), **Taffler & Tisshaw** (1977), **Altman et al.** (1995), and **Lussier** (1995). Nevertheless, most of these authors mentioned in their researches that two basic assumptions of MDA are often broken when applied to the failure prediction.[5] Further discussions on this topic are presented by **Barnes** (1982), **Karels & Prakash** (1987) and **Leay and Omar** (2000). Regarding these issues, the conditional logit model has been employed for the first time by **Ohlson** (1980). By using this methodology, the restrictive assumptions of MDA are not required and the conditional logit model permits dealing with categorized data. The logistic model also returns a score between zero and one, which easily can be converted to the probability of default (PD). Another advantage is that the estimated coefficients can be interpreted one by one as the importance of each of the independent variables in the explanation of the PD. **Wiginton** (1980) in his research found that logit model outperforms discriminant analysis. Following Ohlson and Wiginton, many researchers such as **Campbell and Dietrich** (1983), **Zavgren** (1983), **Gentry** (1985), **Aziz et al.** (1988), **Gardner and Mills** (1989), **Platt et al.** (1990), **Lawrence and Arshadi** (1995), **Becchetti** (2003), **Kleimeier et al.** (2007), **Avery et al.** (2012) also used logit models to predict default. Recently, **Charitou, Neophytou and Charalambous** (2004) demonstrated once again that the logit method outperforms other methods based on their empirical results.

Other statistical techniques have also been suggested to improve the prediction quality of credit scoring models such as Bayesian methods, neutral network, decision tree, k-nearest neighbor, survival analysis, fuzzy rule-based system, support vector machine, and hybrid models. However, the logistic regression still remains the most popular method. Nowadays, even in the largest financial institutions in the world, logistic regression is always considered as one of the main approaches to predict defaults.

Our paper falls into this strand of the literature by providing an empirical analysis of credit scoring in the case of auto loans. From a methodological point of view, the paper focuses mostly on the credit-scoring process using different statistical techniques. From an applied perspective, we provide an application on real data of a France-based retail bank. The behavioral scorecard developed in the paper is a key element of the research. In fact, application scorecards have been investigated in different academic papers whereas behavioral scorecards have not received as much attention by researchers. Additionally, the presentation of real behavioral credit-scoring process is something occasionally encountered, and nothing is done in a form similar to the paper. Finally, it should be noticed that our credit-scoring model has been successfully implemented in the bank which provided the data, hereafter called bank A.

The rest of paper is organized as follows. Section 2 presents the credit scoring development process. Section 3 describes our case study model. Section 4 is devoted to the model application and its results. Section 5 concludes the paper.

2. Credit Scoring Development Process

2.1. Data Preparation & Gathering

Data is of paramount importance to the scorecard development process. Data preparation is time-consuming and consists of the following four main activities:

[4] In the original Z-score model, Altman combined a set of five business ratios: EBIT/Total Assets, Net Sales/Total Assets, Market Value of Equity/Total Liabilities, Working Capital/Total Assets, and Retained Earnings/Total Assets.

[5] The multiple discriminant analysis technique is based on two restrictive assumptions: (i) Multivariate normality of independent variables and (ii) Equal covariance matrices of groups (default/non-default).

- Characteristic identification: Scorecards are based on characteristics, which are in turn translated into questions that the underwriter asks a customer. The questions depend on the type of scorecard, product, market, local compliance and laws. Thus, each characteristic has varying importance and predictive power in the model.

- Selecting the development sample: There are two basic types of data that can be used in the scorecard: demographic data (data collected at the time of application) and performance data (data that covers total historical delinquency performance since time of opening). Different data will be used depending upon the type of scorecard. For example, an application scorecard uses only demographic data captured at the time of application, while a collection/ behavior scorecard is unlikely to use much demographic data, but will rather concentrate on using performance data.

- Checking data quality: After collecting data, the underwriter needs to check data quality. Data need to be complete (contains all relevant information), accurate (being true) and the processing of data needs to be robust (able to withstand changes in the environment).

- Good/ Bad definition: Good/ Bad account definition determines whether accounts are likely to roll to write off or not. For instance, good accounts are likely to repay fully and are those that have always been past due of at most 30 days but no more, whereas bad accounts are those that have been delinquent for 90 days or more. However in every case, it is recommended to carry out the analytical method called "Roll Rate Analysis" in creation of the good/bad definition (**Naeem**, 2006). [6]

2.2. Data transformation: Fine & Coarse Classing

Following the preparation of data, the underwriter has clean data in a form ready to use for modeling. The next step is to analyze the characteristics suitable for being modeled in the data preparation stage. The purpose is to identify the characteristics (or questions) that can separate good accounts from bad ones. Hence, a predictive characteristic contains attributes that display clearly different levels of risk for different attributes. There are two stages in this process: Fine and Coarse Classing.

(i) Fine Classing

In fine classing step, each attribute is broken down and analyzed. Fine classing presents an opportunity to review the data in a summarized form. It is the first step along the road to coarse classing data and identifying characteristics to be included in the scorecard.

At this stage, the underwriter determines if there are any characteristics that have such low predictive power that there is no benefit in including them in any further analysis. To do this, firstly the underwriter calculates the good/bad odds. It is defined as:

$$Odds = \frac{\%Good}{\%Bad}$$

The Odds measure the proportion of good accounts to bad accounts. If this ratio is zero, then there is equal chance of an account with that attribute being good or bad. If it is less than 1 then there are more Bads than Goods.

With the good/bad odds, we can now calculate *weights* for all fine classes. *Weight* is a measure of how good or bad the accounts are within a particular attribute:

$$Weight\ of\ evidence = log(Odds)$$

A weight shows whether there are more Goods than Bads for a given attribute, or vice versa. If the weight is negative, this means there are more Bads than Goods, and if it is positive, it means there are more Goods than Bads. If it is close to zero then there are similar numbers of Bads and Goods.

[6] Roll Rate Analysis helps to predict losses based on delinquency, involves comparing worst delinquency in a specified previous buckets with that in the next buckets, and then computing the percentage of accounts that maintain their worst delinquency, or get better, or roll forward into the next delinquency buckets.

The weights are then used to calculate the *Information Value* (IV) of a characteristic. This is first done for all attributes of a characteristic. These are then totaled to give an information value for the characteristic as a whole.

$$IV = Weight \times (\%Good - \%Bad)$$

The Information Value measures the ability of a characteristic to separate between good and bad accounts. Information Value is always greater than 0 but is generally less than 2.

Using the IV, the underwriter can exclude non-predictive variables. Following is one of the most frequently used rules:

$Exclude: IV < 10\%$

$Satisfactory: 10\% \leq IV \leq 100\%$

$Highly\ predictive: IV > 100\%$

A characteristic with a very low information value is considered to be non-predictive. Calculation of fine classes with information value and exclusion of non-predictive variables is the first step in the process of coarse classing. Obviously, each lending institution has its own policy regarding the exact threshold less than which IVs are considered non-predictive.

(ii) Coarse Classing

The objective of coarse classing is to combine the small fine classes into larger groups so that the number of accounts in each group is significant enough. This optimizes the discriminatory power of attributes and ensures the model is stable. The underwriter should look for similar bad rates as shown by the weights when deciding which groups to put together. This process in reality tends to be manual and requires significant human judgment by the credit officer.

2.3. Modeling Techniques

Once the final coarse classes have been decided, the next step is to model the relationship between the variables, which may be usually either linear or logistical. A linear regression is used for a continuous outcome variable, whereas a logistic regression is retained for a binary outcome variable. Input to regression model comes from coarse classing output.

Scorecards are based on multiple regression which analyzes several variables to predict an outcome.

There are three types of regression: forward selection, backward elimination and stepwise regression. Forward selection starts with an empty model, and then adds a new variable at each step, only stops adding when no further benefit from additional variable. In contrast, backward elimination starts with model containing all possible variables, and then removes a variable at each step, only stops removing when we see deterioration of model performance. Finally, stepwise regression is a combination of these two previous regressions.

Generally, the underwriter should use 5% or 10% significance level but also consider adjusting the significance level to remove/add variables so that the number of variables should be as relevant as possible.

2.4. Checking for Correlation

The underwriter needs to check the correlation between every pair of variables in our model, and considers removing one of the variables if correlation is higher than 0.5 in absolute value (i.e. if the two variables are “too” correlated). The choice of the rule is not rigid and can vary according to the modeler’s personal opinion if this preserves the model’s stability and performance.

2.5. Analyzing Model Performance

Score validation provides a measure of the performance of the scorecard. There are several different measures that can be used to assess the effectiveness of a scorecard before it is implemented, as well as during the monitoring phase. These measures are summarized in Table 1.

Table 1: Principal Performance Measurements with Decision-making Rules

1. Kolmogorov-Smirnov statistic	**2. Gini coefficient**	**3. Score distribution**	**4. Bad rate and Good/ Bad Odds by Score**
The KS measures the widest spread between cumulative Goods and cumulative Bads. The divergence between the two curves determines the strength or weakness of the scorecard to differentiate good customers from bad customers. In other words, the higher KS, the better the model since Goods are more separated from Bads. Calculation: If we divide the sample into 10 classes by descending risk level, the cumulative percentage of Bads in the i^{th} class is noted cp_{B_i}. Similarly, the cumulative percentage of Goods in the i^{th} class is noted cp_{G_i}. The calculation of the KS indicator is presented as follows: $D_{KS} = Max_i\{\lvert cp_{B_i} - cp_{G_i}\rvert\}$	Like KS, the Gini coefficient is a quantitative measure of how well the model discriminates between Goods and Bads, but by looking at actual discrimination versus perfect discrimination. Like KS, the calculation of the Gini is presented below: $D_{Gini} = \sum_{i=1}^{10}\{(cp_{G_i} - cp_{G_{i-1}})(cp_{B_i} + cp_{B_{i-1}})\} - 1$	Score distribution is another measurement that the underwriter looks for a well-distributed plot.	This measurement is necessary to verify if the bad rate is decreasing as the score increases and the good/bad odds increase as the score rises.
A good KS is normally superior to **35**% and higher is better. A higher KS value indicates that the scorecard is doing a better job of separating Goods from Bads, and thus, rank-ordering risk.	A Gini coefficient higher than **40**% is considered to be satisfactory. The higher the Gini, the better the model since again, Goods are better separated from Bads.	The ideal score distribution is the one with no score having more than **5**% of accounts in it.	If the rule is violated, the scorecard will make no sense.

2.6. Implementation and Use

The implementation analysis starts by comparing the performance of the new scorecard to that of the existing one. If the new scorecard cannot discriminate better between Goods and Bads then it is unlikely that it would be implemented. The underwriter should compare the scorecards by looking at their respective KS and Gini statistics.

A scorecard is built upon a data sample taken from the past, and the performance of the scorecard is therefore reliant upon the assumption that the past will accurately represent the future. Since the past and the future normally will not be the same, it is important that the underwriter monitors the portfolio to identify and incorporate any differences into the scorecards as quickly as possible.

Monitoring is also an important part in understanding the portfolio and how it is performing, and ensuring that the scorecard is making optimal decisions given the available information.

This section outlines the entire credit scoring process and also demonstrates that credit scoring is not simply a mathematical model of predicting defaults. In reality, it requires an important degree of human judgment during its development and implementation process. In fact, by finding considerable differences between the implied loss distributions of the two banks with equal regulatory risk profiles,

Jacobson et al. (2006) prove that the formal design of a rating system and the way in which it is implemented can be quantitatively important for the shape of credit loss distributions. Moreover, if the loan officers use unique data based on hard information to develop credit scoring models, and they are volume-incentivized, **Puri et al.** (2013) show that loan officers increasingly use multiple trials to move loans over the cut-off, both in a regression-discontinuity design and when the cut-off changes. Another relevant research to address the incentives of financial intermediaries is presented by **Keys et al.** (2010). Their empirical result proposes that securitization, by converting illiquid loans into liquid ones, could reduce lenders' incentives to screen borrowers, thus limiting the utility of credit scoring.

3. Model Description

The behavioral credit-scoring model presented in this paper is used, for the first time, to estimate the probability of default of all individual automobile loans in bank A, based mostly upon the customers' financial, family standing and payment behaviors. This scorecard will be applied for each current loan in the individual automobile loans portfolio so as to assess the quality of the portfolio.

The model is developed in the SAS language, using the assumption that future performance will reflect past one. The model is then expected to be stable over time. However, substantial changes in macroeconomic environment, government regulation, product specifications, and population might affect scorecard stability.

Variables used in this scorecard include both application and behavior data. Application data presents some of the following characteristics: employment seniority, house ownership, marital status, family income, rent amount, and other comprehensive expense amount. Behavioral data includes worst delinquency in last three months, number of month-in-default, time since last delinquency, and other payment behavior variables. Hence, all data comes from loan application forms, collected during the lifecycle of customers. There is no use of external data, like credit bureau data or market information.

All available data from the observation window was used for the development sample (80% of outcome periods) and validated in the validation sample (20% remaining outcome periods). Three observation dates have been chosen: 01/2010, 04/2010, and 07/2010; and three outcome periods have been considered:

- February 2010 to January 2012
- May 2010 to April 2012
- August 2010 to July 2012

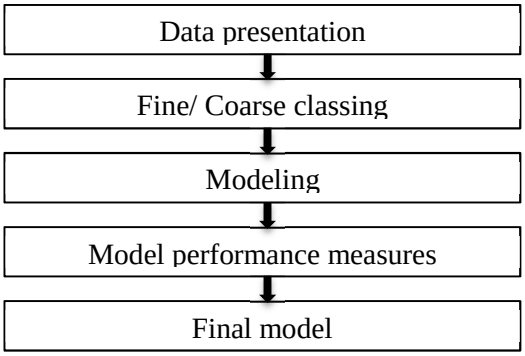
This model has been implemented since 01/07/2013 after a validation performed by an independent model validation unit of the bank.

4. Application and Results

The model development begins with data gathering and processing. This phase aims to understand the data and its quality in order to define the scope of study.

The scorecard development process is summarized in Figure 1.

Figure 1: Scorecard development process



4.1. Data preparation

The model depends totally upon data, and there are many data considerations during the scorecard development and use. Data is then verified to make sure of its transparency, its quantity and its quality.

Application and behavior data

As previously mentioned, application data are those collected at time of application such as marital status, residential status, age, income, and other comprehensive expense amount. In addition, some variables included in application data may not be used in their raw format but used to generate other variables. The following are example of generated variables:

- Time with bank = Date of situation – Date of first account opening (in years)
- Time at present address = Date of situation – Date of entry to the house (in years)

Behavioral data are essential to have an overall look at the history of delinquency performance since time of opening, in order to predict future clients' payment behaviors and portfolio quality.

The following are examples of behavioral generated variables:

- Maximum bucket (or number of days past due) in the last 12, 6 and 3 months
- Number of unpaid months in the last 12, 6 and 3 months

Generated cross-variables

These variables are created for the purpose of having new variables with a stronger information value. Examples of cross-variables include age-revenue, loan term-loan amount.

A total of 32 variables are initially selected for inclusion in the behavioral scoring model. These variables are listed in the first column of Table 2 and divided into five groups: (i) customer's personal characteristics; (ii) loan characteristics; (iii) financial situation; (iv) other behavioral variables; and (v) generated cross-variables. Compared to the currently existing credit scoring models, our list of variables overlaps extensively with **Crook et al.**'s (1992), **Vigano**'s (1993), **Kleimeier et al.**'s (2007), as well as **Kocenda et al.**'s (2011) list of commonly candidate variables (as shown in Table 2).

Table 2: Comparison of candidate variables commonly used in credit scoring models through different researches

Nguyen (2014) - France	Kocenda and Vojtek (2011) - Czech	Kleimeier et al. (2007) - Vietnam	Schreiner (2004) - Bolivia	Vigano (1993) – Burkina Faso	Crook et al. (1992) - UK
Group 1: Customer's personal characteristics (marital status, age, region, residential status, job, time at present address, number of children, time at present employment, time at bank) Group 2: Loan characteristics (type of credit, vehicle type, vehicle condition, vehicle price, loan amount, loan duration, client's contribution-to-vehicle price ratio, client's initial contribution, existence of a co-borrower) Group 3: Financial situation (income, debt-to-income ratio, troubled debt restructuring) Group 4: Other behavioral variables (number of months in bucket 0, max bucket in the past 12/6/3 months, number of months of non-payment in the past 12/6/3 months, delinquency status) Group 5: Generated cross-variables (age-revenue, loan term-loan amount, marital status-revenue)	Socio-demographic variables: Education (1), Marital status (2), Years of employment (3), Sector of employment (4), Gender (5), Date of Birth (6), Type of employment (7), Number of employments (8), Employment position (9), Credit ratio 1 (10), Credit ratio 2 (11), Region (12) Bank-client relationship variables: Own resources (13), Amount of loan (14), Purpose of loan (15), Length of client/bank relationship (16), Date of account opening (17), Deposit Behavior (18), Loan Protection (19), Type of product (20), Number of co-singers (21), Date of loan (22).	Age (1) Gender (2) Marital status (3) Education (4) Residential status (5) Time at present address (6) Monthly income (7) Occupation (8) Time with present employer (in years) (9) Loan purpose (10) Loan duration (11) Collateral to loan ratio (12) Time with bank (13) Number of prior loans (14) Number of current accounts (15) Number of saving accounts (16) Region (17)	Date of disbarment (1) Amount disbursed (2) Type of guarantee (3) Branch (4) Loan officer (5) Gender of the borrower (6) Sector of the firm (7) Number of spells or arrears (8) Length of the longest spell of arrears (9)	Group 1: Customer's personal characteristics (age, sex, religion, marital status, education, employment sector and place, etc.) Group 2: Data on the enterprise (type, professional skills, number of employees, productivity, profitability, etc.) Group 3: Profitability (main and secondary revenue, revenue stability) Group 4: Amount and composition of assets (total assets including money and deposits) Group 5: Financial situation (initial and current amount of loans received, defaults, loans granted) Group 6: Investment plans (presence of investment plan, other sources of finance) Group 7: Customer's relationship with bank (past loans with bank, savings account with bank, etc.) Group 8: Bank's control of credit risk (loan destination, disbursement form, method collateral, contractual conditions on interest rates, etc.)	Postcode (1) Employment status (2) Years at bank (3) Current account (4) Spouse's income (5) Residential status (6) Phone (7) Years at present employment (8) Deposit account (9) Value of home (10) Outgoings (11) Number of children (12) Applicant's income (13) Mortgage balance outstanding (14) Charge card (15)

4.2. Fine & Coarse classing

The fine and coarse classing step needs to be done before modeling the score. This step aims at building homogeneous classes in terms of risk or bad rate, and also identifying the characteristics that can better separate the Goods out from the Bads.

For each variable, we apply the "IV" (Information Value) method:

$$\boldsymbol{Weight\ of\ Evidence\ of\ characteristic\ i} = \boldsymbol{log}(\boldsymbol{Odds\ ratio\ of\ i})$$

$$= \boldsymbol{log}\left(\frac{\%\boldsymbol{G_i}}{\%\boldsymbol{B_i}}\right)$$

$$= \boldsymbol{log}\left(\frac{\boldsymbol{G_i}}{\boldsymbol{B_i}} \times \frac{\sum \boldsymbol{B}}{\sum \boldsymbol{G}}\right)$$

$$\boldsymbol{IV} = \boldsymbol{Weight} \times (\%\boldsymbol{G_i} - \%\boldsymbol{B_i})$$

This method is first applied to continuous variables and then to categorical variables. We categorize continuous variables as suggested by **Thomas et al.** (2002). First, the range of values for each continuous variable was split into ten categories, based on the assumption that all categories should have the same number of observations. Second, odds ratios and information values were computed for each category (fine classing) and categories with similar values were joined together (coarse classing). This step was also performed for categorical variables.

Let us now present a number of examples for both types of variables.

Example of a continuous variable: loan term

A common approach for the initial fine classing of a continuous variable is to create a standard number of equally sized groups. In this case, the variable is sorted and classed into at most 10 intervals. Tables 3 and 4, as well as Figures 2 and 3 summarize the results obtained for loan term.

Table 3: Fine classing of loan term

Loan term (months)	%Good	%Bad	Odds	Weight	IV	Bad Rate
0 – 35	3.6%	2.0%	1.8	0.6	1.0%	4.0%
36- 48	19.9%	15.5%	1.3	0.2	1.1%	4.1%
49 - 59	6.4%	6.1%	1.1	0.1	0.0%	5.8%
60 - 72	68.1%	75.6%	0.9	-0.1	0.8%	5.8%
>=73	2.0%	0.8%	2.4	0.9	1.0%	2.3%
Total					**3.9%**	

Table 4: Coarse classing of loan term

Loan term	%Good	%Bad	Odds	Weight	IV	Bad Rate
<= 48	23.5%	17.5%	1.3	0.3	1.8%	4.0%
49 - 72	74.5%	81.7%	0.9	-0.1	0.7%	5.8%
> 72	2.0%	0.8%	2.4	0.9	1.0%	2.3%
Total					**3.5%**	

Figure 2: Fine Classing of loan term

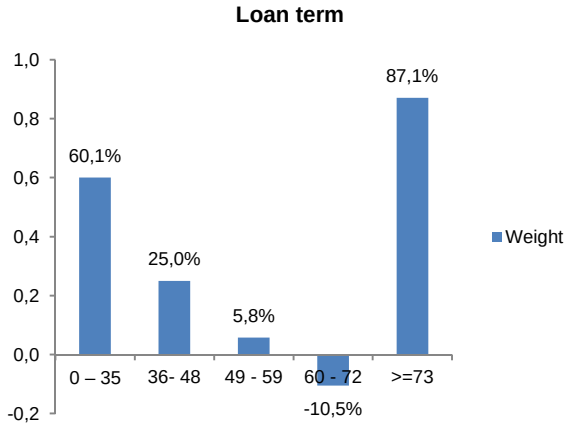


Figure 3: Coarse Classing of loan term

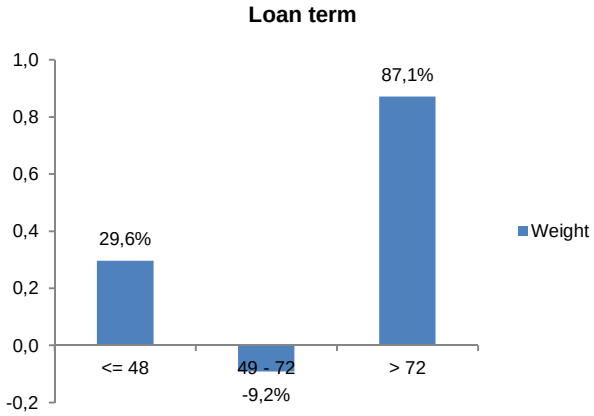


Figure 2 displays the results of the fine classing step: the population is distributed into five classes. A negative log odds (weight) means that the class is more risky than the average. At this stage, the IV sum gives a discriminating power of **3.9**%.

In the coarse classing step (Table 4), we look at the similar bad rates, and then create three homogeneous classes for the variable. Coarse classing gives us more visibility even if we lose a little bit of significance.

Example of a categorical variable: Residential status

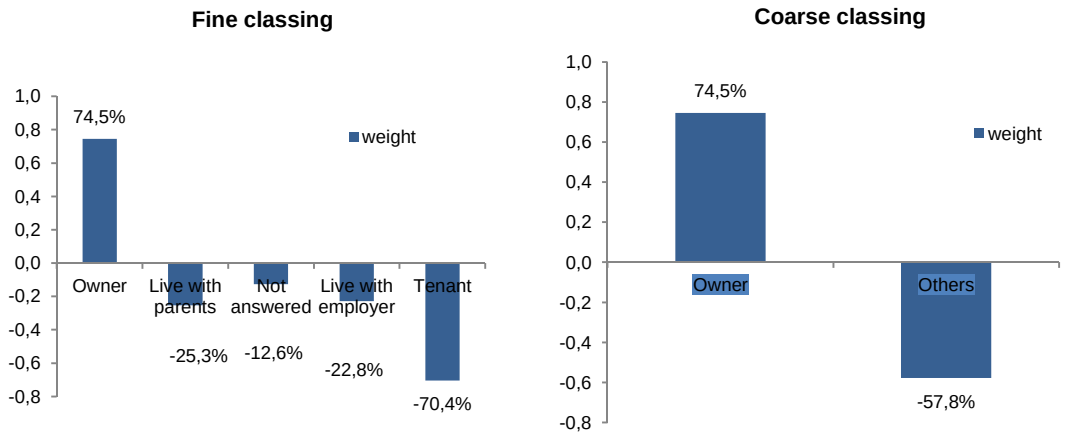
Table 5: Fine classing of residential status

Residential status	%Good	%Bad	Odds	Weight	IV	Bad Rate
Owner	59.8%	28.4%	2.1	0.7	23.4%	2.9%
Live with parents	10.7%	13.8%	0.8	-0.3	0.8%	7.4%
No answer	0.2%	0.2%	0.9	-0.1	0.0%	6.8%
Live with employer	2.1%	2.6%	0.8	-0.2	0.1%	7.2%
Tenants	27.2%	55.0%	0.5	-0.7	19.6%	9.9%
Total					**43.9%**	

Table 6: Coarse classing of residential status

Residential status	%Good	%Bad	Odds	Weight	IV	Bad Rate
Owner	59.8%	28.4%	2.1	0.7	23.4%	2.9%
Others	40.2%	71.6%	0.6	-0.6	18.2%	8.4%
Total					**41.6%**	

Figure 4: Fine and Coarse Classing of Residential status



Tables 5 and 6 and Figure 4 summarize the analysis of residential status within the behavioral scorecard development. Once fine classing is done, the next stage is to combine the small fine classes into larger groups so that the number of accounts in each group is significant enough.

However, there are two opposite purposes when implementing coarse classing. On the one hand, it is better to represent the characteristic with as few coarse classes as possible. The reason is the smaller the number of classes, the less complicated and over-fitting the model. On the other hand, having too few coarse classes can make lose valuable information because the smaller the number of potential coarse classes, the less powerful the variable, and vice versa.

Thus, for residential status, the goal is to reduce the number of classes as much as possible, but with a minimal information loss. In this case, merging different fine classes into two larger coarse classes, which are owner and others, reduces the IV from **43.9**% to **41.6**%.

The information values of all candidate variables can be found in Table 7.

Table 7: List of candidate variables with Information Value

Variables	Fine classing		Coarse classing	
	IV > 10 %	**2% < IV <= 10 %**	**IV > 10 %**	**2% < IV <= 10 %**
Marital status	15,58%		12,27%	
Residential status	43,9%		41,6%	
Type of credit		2,37%		
Vehicle condition		1,15%		
Vehicle type		2,04%		
Region		6,19%		5,92%
Existence of co-borrower(s)		1,77%		
Troubled Debt Restructuring	45,99%			
Job	20,41%		19,01%	
Loan duration		3,9%		3,5%
Time at present address	11,17%			9,52%
Loan amount	23,48%		14,80%	
Income	13,88%		12,97%	
Number of children		2,30%		2,29%
Time at present employment	18,39%		14,91%	
Age of the client	13,93%		12,82%	
Client's Initial contribution	41,40%		37,39%	
Client's contribution-to-vehicle price ratio	48,05%		43,56%	

Debt-to-income ratio	27,58%		23,53%	
Time at bank		0,04%		
Price of vehicle		3,86%		3,17%
Number of months in Bucket 0	193,94%			
Max bucket in the past 12 months	206,27%		201,19%	
Max bucket in the past 6 months	193,55%		189,02%	
Max bucket in the past 3 months	180,97%		176,82%	
Months of non-payment past 12 months	199,69%		196,21%	
Months of non-payment past 6 months	187,44%		182,92%	
Months of non-payment past 3 months	173,51%			
Delinquency status	105,16%			
Age-Revenue	16,08%		15,37%	
Loan term - Loan amount	18,92%		17,79%	
Marital status - Revenue	16,64%		15,21%	

Obviously, the most significant variables are all behavioral ones. They are those that characterize historical bank-client relationships, a finding that is in line with the comprehensive overview in **Anderson** (2007). The remaining variables are significant for the most parts.

4.3. Modeling

Before entering in the modeling step, it is necessary to review characteristics for possible inclusion in a scorecard. There are some primary factors to be considered:

- The variables need to be relevant.
- They have a significant degree of predictive power (through their Information Value).
- They have a low correlation with each other (through Cramer's V & Pearson's test). [7]

At this step, only the variables that are adapted to these criteria for modeling are selected. Next, to model the relationship between the variables, the logistic regression method works best for binary criterion (Good/Bad) and is recommended by bank A's global banking policy. Such regression is estimated using the Maximum Log-likelihood method.

Logistic regression requires some assumptions as follows: (i) categorical target variable; (ii) linear relationship, with the log odds functions; (iii) independent error terms; (iv) non-correlated predictors; and (v) use of relevant variables.

We have tested 8 models, each with a different combination of variables. To this end, we rely on automated selection routines, such as forward selection, backward elimination and stepwise regression.

Compared to the other two techniques, the backward elimination method has notable advantages. First, a set of variables could have a significantly higher predictive power even if any subset of them does not. This is a default of forward selection and stepwise regression, which do not admit a variable into the model simply because its individual predictive power is not good enough. Furthermore, backward elimination measures the model's joint predictive power as a whole because it starts with all of the variables. Thus, the backward method is selected.

4.4. Model Performance Measures

[7] Cramer's V shows the degree of association in tables which have more than 2x2 rows and columns. It is calculated as $V = \sqrt{\frac{\chi^2}{N(k-1)}}$, where χ^2 is derived from Pearson's test, N is the the total number of observations and k is the number of rows or columns in the table. Cramer's V may vary between 0 and 1. A value close to 0 indicates little association between variables and vice versa.

Once the modeling step is done, we assess the model performance. At this stage, all candidate models need to be passed through performance tests.

First, the candidate model is rejected if one of the following criteria is violated:

i. KS is generally higher than 35%.
ii. Gini is normally superior to 40%.
iii. Regarding to score distribution, no score has more than 5% of accounts.
iv. The bad rate decreases when the score increases and the good/bad odds increase when the score increase.
v. There is no significant change (ideally less than 1%) in the criteria values above of the development sample and ones of the validation sample.

Second, we keep only models that satisfy the criteria, and then choose the best model by comparing the criteria values of different models.

Finally, we need to make sure that the final model can validate every performance criterion.

As a result, we have tested 8 models, following the three steps presented above, and have finally found one model which outperforms the others. The final model contains 10 variables which are presented in Table 8.

Table 8: List of variables in the selected behavioral scoring model

1. Marital status
2. Residential status
3. Troubled debt restructuring
4. Job
5. Client's contribution-to-vehicle price ratio
6. Debt-to-income ratio
7. Number of months in the bucket 0
8. Max bucket in the past 6 months
9. Months of non-payment in the past 6 months
10. Delinquency status

We provide below performance measurement results of our best model.

KS & Gini

Table 9 and Figures 5 to 7 display the results regarding KS statistic and Gini coefficient.

Table 9: Comparison of KS and Gini in the development and validation sample

Development sample		**Validation sample**	
KS	**Gini**	**KS**	**Gini**
58.94%	74.64%	58.74%	74.39%

Figure 5: KS of the development sample

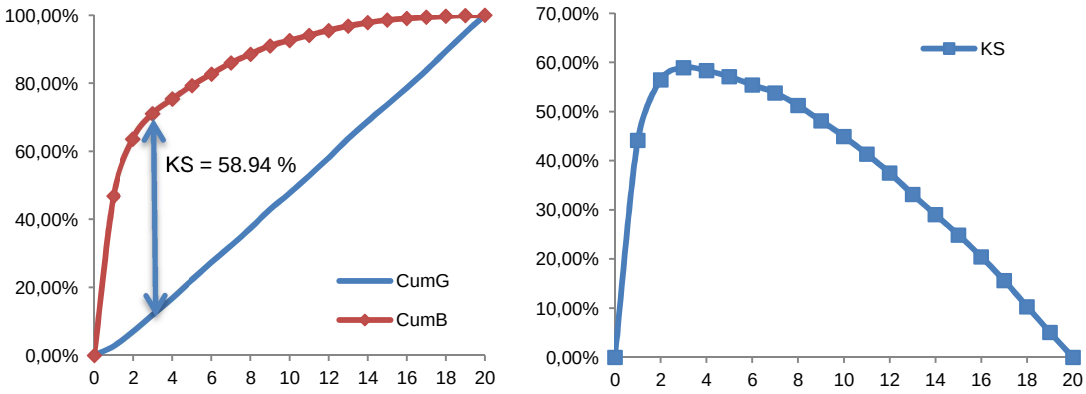
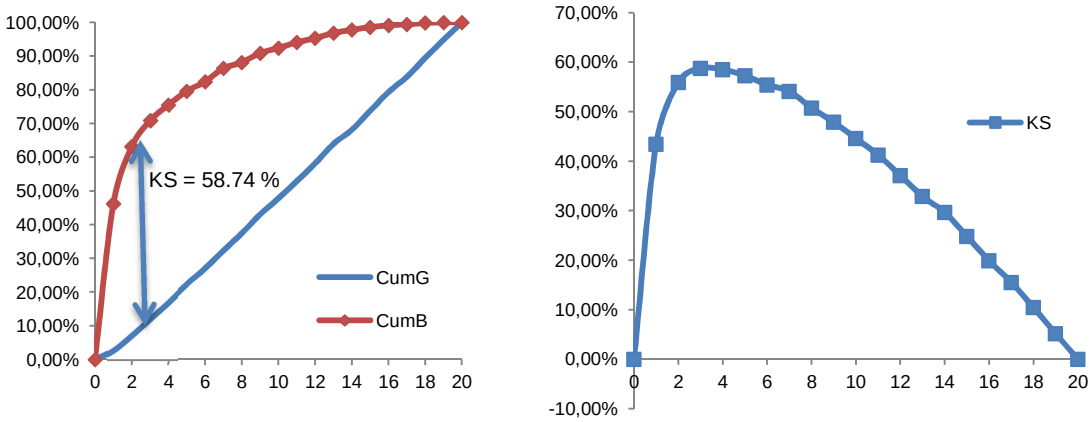
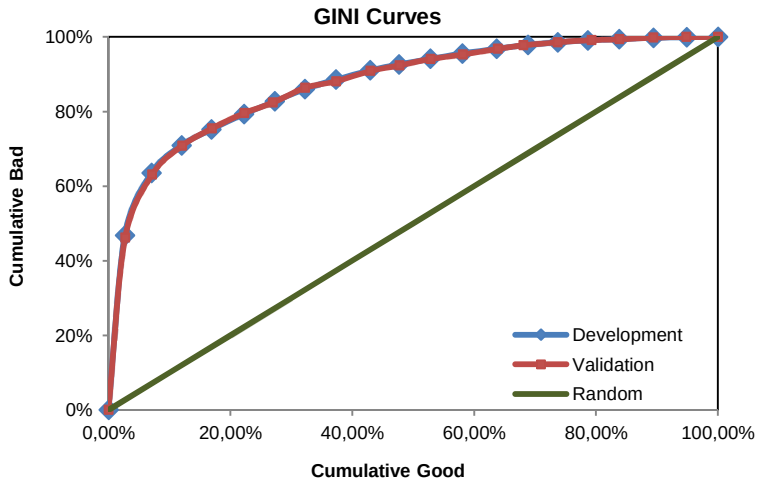


Figure 6: KS of the validation sample



As shown, we observe a small variation in KS and Gini in the development and validation samples. KS and Gini are also greater than 50%. These elements confirm the power and the stability of the model in the whole population.

Figure 7: Gini Curves Comparison



The Lorenz curve (Figure 7) shows the cumulative percentage of good against the cumulative percentage of bad in the development and validation samples. The green line indicates the distribution of good and bad under random scoring where every customer has a 50-50 chance of being good (or bad), whereas the red (validation sample) and blue (development sample) lines show the plot using our

model in each of the samples. Also, there is only a nearly imperceptible difference between the Lorenz curve of the development sample and that of the validation sample. The performance of the model is thus confirmed.

Bad rate and Good/ bad odds by score

Figure 8: Bad Rate and Odds in the development and validation samples

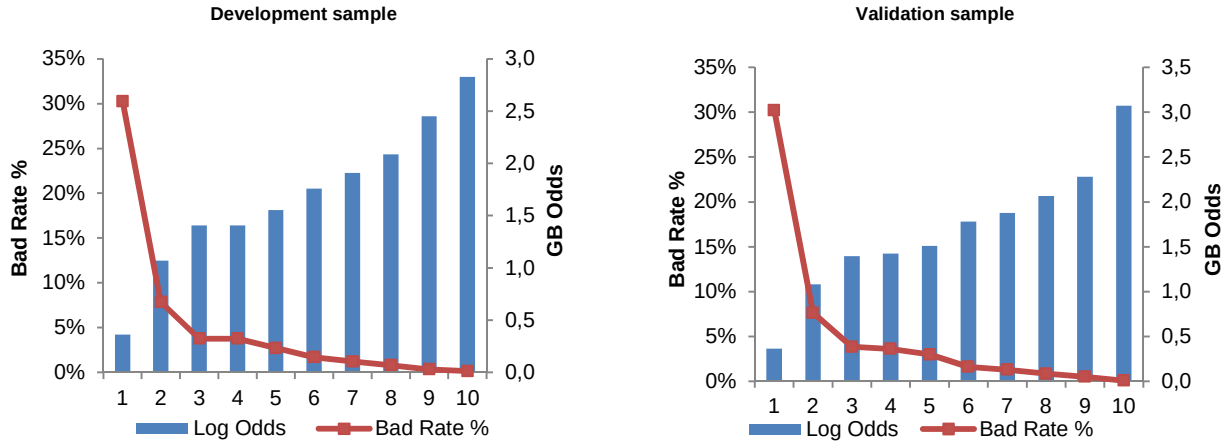
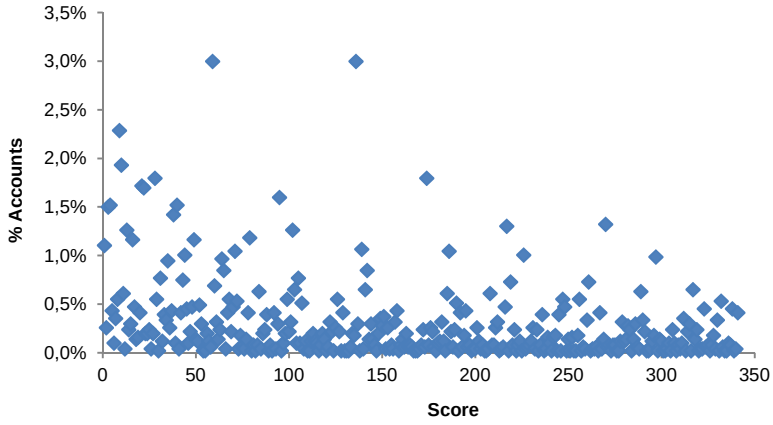


Figure 8 shows that the bad rate decreases as the score increases and the good/bad odds increase as the score increases. This is obviously a desirable characteristic of the model.

Score Distribution

Figure 9: Score distribution



The score distribution (Figure 9) shows our expected result. This is a well-distributed plot with no score having more than 4% of accounts. In other words, the scores given by the model is rather homogeneous and not concentrated.

All these results confirm the performance of our selected model.

4.5. Calibration & Implementation

Calibration

Score must now be calibrated and presented in the proper unit. In other words, calibration is a way of aligning our scorecard so that the odds by score relationship are consistent across all scorecards.

Let f_{target} be the calibrated (target) score. The alignment process is simply given by:

$$f_{target} = \alpha_0 + \alpha_1 \times Score$$

As the result of applying to our specific model, score calibration is presented as follows:

$$\boldsymbol{FinalScore = 419.2 + 1.5 \times LocalBehaviorScore}$$

The *FinalScore* value represents the final score that is consistent across all scorecards of the bank, regardless of the country where those scorecards are established.

Finally, we can sort the population based on this metric. Five credit ratings are proposed to tally the population into groups of varying credit risk degrees. The result is shown in Table 10:

Table 10: Credit Rating

Credit Rating	Condition	Frequency
A+	$FinalScore \geq 750$	33,35%
A	$700 \leq FinalScore < 750$	28,92%
B	$650 \leq FinalScore < 700$	24,39%
C	$600 \leq FinalScore < 650$	7,62%
D	$FinalScore < 600$	5,72%

Implementation

Before a model can be put into use, an independent model-validation unit must document if model validation results are correct. Internal audit will then verify that each model enters production with formal approval by the validation unit.

The validated model will be now used to assess the portfolio's quality. To achieve this, we must first apply the final equation to score every present accounts. These observations after calibration will be tallied into five credit ratings (CR) as shown above, as well as into a number of groups based on the accounts' default probability (PD). The expected gross value of losses can be estimated by the following relationship:

$$Gross\ Expected\ Loss = \sum_{i=1}^{N} PD_i \times Outstanding\ Amount_i$$

The expected net value of losses is calculated by taking into account the recovery rate:

$$Net\ Expected\ Loss = Gross\ Expected\ Loss \times (1 - Recovery\ Rate)$$

These measures allow us to calculate the provision amount corresponding to the cost of risk of the portfolio. One can therefore determine the evolution of the portfolio's risk and analyze its quality.

5. Conclusion

The objective of the paper is to show how scorecards are built within a lending institution by giving a detailed development process of a real behavioral scorecard. We have provided a glimpse of the credit risk management process and how loans are classified, what are the data needed for scoring and how they are selected, and finally how the models are implemented, checked for validity and compared between themselves. Our paper is mostly in line with existing researches, and has the additional advantage of describing how credit scoring is applied in real life.

This paper may be extended for further research. First, it would be interesting to compare the result of the logistic regression with the results of other statistical techniques. Second, we should extend the current credit process to collection/fraud by developing a collection/fraud scorecard for the purpose of being more efficient in optimizing the net expected loss and then improving the portfolio management process. Third, analyzing the advantages/defaults of different performance indicators (such as KS, Gini, Score distribution, etc.) used to assess scorecard power would be a promising extension.

References

ALPERT M. and RAIFFA H. (1969), *A Progress Report on the Training of Probability Assessors*, Unpublished manuscript, Harvard University. In: KAHNEMAN D., SLOVIC P., and TVERSKY A. (1982, eds), *Judgment under Uncertainty: Heuristics and Biases*, 294–305, Cambridge University Press: New York.

ALTMAN E.I. (1968), *Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy*, Journal of Finance 23(4), 589-611.

ALTMAN E.I., HALDEMAN R.G., and NARAYANAN P. (1977), *Zeta-analysis: A New Model to Identify Bankruptcy Risk of Corporations*, Journal of Banking and Finance 1, 29-54.

ALTMAN E.I., HARTZELL J. , and PECK M. (1995), *A Scoring System for Emerging Market Corporate Debt*, Salomon Brothers Emerging Markets Bond research.

ALTMAN E. I. and SAUNDERS A. (1998), *Credit Risk Measurement: Developments over the last 20 years*, Journal of Banking and Finance, Vol.21, pp.1721-1742.

ANDERSON R. (2007), *The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation*, Oxford University Press.

AVERY R.B., BREVOORT K.P. and CANNER G. (2012), *Does Credit Scoring Produce a Disparate Impact?*, Real Estate Economics, American Real Estate and Urban Economics Association, vol. 40, pages S65-S114, December.

AZIZ A., EMANUEL D.C., and LAWSON G.H. (1988), *Bankruptcy prediction - An investigation of cash flow based models*, Journal of Management Studies 25 (5), 419-437.

BARNES P. (1982), *Methodological implications of non-normality distributed financial ratios*, Journal of Business Finance and Accounting 9(1), 51-62.

BARBER B. and ODEAN T. (1999), *The Courage of Misguided Convictions: The Trading Behaviour of Individual Investors*, Financial Analysts Journal 55 (6), November/December, 41–55.

BEAVER W. (1967), *Financial ratios predictors of failure*, Journal of Accounting Research 4, pp. 71-111.

BECCHETTI L. and SIERRA J. (2003), *Bankruptcy risk and productive efficiency in manufacturing firms*, Journal of Banking and Finance, 27 (11), pp. 2099-2120.

BIERMAN H. and HAUSMAN W. H. (1970), *The credit granting decision*, Management Science, April, v 16, 519-32.

BLUM M. (1974), *Failing company discriminant analysis*, Journal of Accounting Research, 12(1), 1-25.

CAMPBELL T. S. and DIETRICH J. K. (1983), *The Determinants of Default on Insured Conventional Residential Mortgage Loans*, Journal of Finance, vol. 38, 1983, pp. 1569-1581.

CHARITOU A., NEOPHYTOU E., and CHARALAMBOUS C. (2004), *Predicting Corporate Failure: Empirical Evidence for the UK*, European Accounting Review, vol. 13, 2004, pp. 465- 497.

CROOK J.N., HAMILTON R. and THOMAS L.C. (1992), *A comparison of discriminations under alternative definitions of credit default*, in: Thomas, L.C., Crook, J.N., Edelman, D.B. (Eds.), *Credit Scoring and Credit Control*, Oxford University Press, Oxford, pp. 217- 245.

DEAKIN E. (1972), *A discriminant analysis of predictors of business failure*, Journal of Accounting Research, 10(1), 167-179.

DIRICKX Y., and WAKEMAN L. (1976), *An extension of the Bierman–Hausman model for credit granting*, Management Science 22:1229–37.

EDMISTER R. (1972), *An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction*, Journal of Financial and Quantitative Analysis 7(2), 1477-1493.

FALKENSTEIN E., BORAL A. and CARTY V. (2000), *RiskCalcTM for Private Companies: Moody's Default Model Rating Methodology*, Moody's Investors Service, Global Credit Research.

GARDNER, M. J. and MILLS, D. L. (1989), *Evaluating the Likelihood of Default on Delinquency Loans*, Financial Management, vol. 18, 1989, pp. 55-63.

GENTRY J.A., NEWBOLD P. and WHITFORD D.T. (1985), *Classifying bankrupt firms with funds flow components*, Journal of Accounting Research 23(1), 146-160.

HAND D.J. (1997), *Construction and Assessment of Classification Rules*, John Wiley, Chichester, U.K.

HAND D.J. (2001), *Modelling consumer credit risk*, IMA Journal of Management Mathematics 12(1) 139-155.

JACOBSON T., LINDE J. and ROSZBACH K. (2006), *Internal Ratings Systems, Implied Credit Risk and the Consistency of Banks' Risk Classification Policies*, Journal of Banking and Finance 30 (7), 1899-1926.

KARELS G.V. and PRAKASH A.J. (1987), *Multivariate normality and forecasting of business bankruptcy*, Journal of Business Finance & Accounting, 14(4), 573-593.

KEYS B.J., MUKHERJEE T., SERU A., and VIG V. (2010), *Did securitization lead to lax screening? Evidence from subprime loans*, The Quarterly Journal of Economics 125 (1), 307-362.

KLEIMEIER S., and DINH T.T.H (2007), *Credit Scoring for Vietnam's Retail Banking Market*, International Review of Financial Analysis, Elsevier, vol. 16(5), pages 471-495.

KOCENDA E., VOJTEK M. (2011), *Default Predictors in Retail Credit Scoring: Evidence from Czech Banking Data*, Emerging Markets Finance and Trade, 47(6), 80–98.

LAWRENCE E. and ARSHADI N. (1995), *A Multinomial Logit Analysis of Problem Loan Resolution Choices in Banking*, Journal of Money, Credit and Banking, vol. 27, 1995, pp. 202- 216.

LIBBY R. (1975), *Accounting Ratios and the Prediction of Failure: Some Behavioral Evidence*, Journal of Accounting Research, 150–161.

LUSSIER R. N. (1995), *A non-financial business success versus failure prediction model for young firms*, Journal of Small Business Management 33(1), 8-20.

MC LEAY S. and OMAR A. (2000), *The sensitivity of prediction models to the nonnormality of bounded an unbounded financial ratios*, British Accounting Review 32, 213-230.

MEEHL P.E. (1954), *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, University of Minnesota Press: Minneapolis, MN.

NAEEM S. (2006), *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, John Wiley & Sons.

NISBET R., KRANTZ D., JEPSON C., and FONG G. (1982), *Improving Inductive Inference*. In: KAHNEMAN D., SLOVIC P., and TVERSKY A. (eds), *Judgment Under Uncertainty: Heuristics and Biases*, 445–462. Cambridge University Press: Cambridge.

OHLSON J. (1980), *Financial ratios and the probabilistic prediction of bankruptcy*, Journal of Accounting Research 18(1), 109-131.

PLATT H.D. and PLATT M.B. (1990), *Development of a class of stable predictive variables: the case of bankruptcy prediction*, Journal of Business Finance & Accounting 17(1), 31- 51.

PURI M., ROCHOLL J., and BERG T. (2013), Loan officer incentives and the limits of hard information, NBER Working Paper No. 19051

SCHREINER M. (2004), *Scoring arrears at a microlender in Bolivia*, Journal of Microfinance 6, 65–88.

SRINIVASAN V. and KIM Y. H. (1987), *The Bierman-Hausman Credit Granting Model: A Note*, Management Science, 33, 1361-1362.

TAFFLER R.J. and TISSHAW H. (1977), *Going, Going, Gone - Four Factors Which Predict*, Accountancy 88(1083), 50-54.

THOMAS L.C., EDELMAN DAVID B. and CROOK JONATHAN N. (2002), *Credit Scoring and its Applications*, Philadelphia, USA, SIAM.

VIGANO L. (1993), *A credit scoring model for development banks: An African case study*, Savings and Development 4, 441-482.

WIGINTON J.C. (1980), *A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior*, Journal of Financial and Quantitative Analysis, vol. 15, 1980, pp. 757-770.

ZAVGREN C. (1983), *The prediction of corporate failure: the state of the art*, Journal of Accounting Literature 2, 1-37.