

Evaluating the Likelihood of Using Linear Discriminant Analysis as A Commercial Bank Card Owners Credit Scoring Model

John Mylonakis

10 Nikiforou str., Glyfada, 16675, Athens, Greece

E-mail: imylonakis@vodafone.net.gr

George Diacogiannis

University of Piraeus

Department of Banking & Financial Management

80 Karaoli and Dimitriou str., 185 34 Piraeus, Greece

E-mail: gdiak@unipi.gr

Abstract

The paper attempts to determine whether there exists a relationship between the on-time payments of credit card owners of a Commercial Bank and their demographic characteristics (particular personal and family status). It evaluates the statistical technique of discriminant analysis on credit card customers' data of a Greek Commercial Bank and examine whether it is possible to create a model evaluating the credibility of prospective credit-card customer. The sample includes personal data, as well as, payment consistency for 829 customers of the Greek Commercial Bank (X-BANK) of average size. The statistical analysis of the sample data included the identification of the relationship between the theoretical and empirical prices of the distributions of the bank customers' specific variables and discriminant analysis. The results showed that establishing a model to evaluate the credibility of prospective bank card customers, using the technique of the linear discriminant analysis, is not possible. The findings prove interesting and useful for all bank managers. The paper contributes to the financial services literature by adding a further critical analysis into credit scoring systems established by several banking institutions.

Keywords: Bank Credit Card, Credit Scoring Systems, Financial Risk, Bank Consumer Credit Risk, Banking Institutions

1. Introduction

Today, more than ever before in the history of world economy, the use of bank credit cards are extremely high in number and accessible to the largest part of the population of developed countries. Therefore, many Banking Institutions searched for setting up credible evaluation systems (credit analysis, credit scoring systems) in order to facilitate their managers' decisions to accept or reject a new applicant for credit (Nevin and Churchill, 1979; Capon, 1982). Credit scoring is a method of evaluating the credit risk of applicants and predicting their future bank consumer behaviour whether they will default or become delinquent (Mester, 1997). Credit risk forms the primary source of threat for banks. Yet, practical experience shows that scoring systems are not characterized by high efficiency in the long run, as they are based on borrowers' qualitative characteristics that cannot be adequately quantified. Furthermore, that credit scoring raises significant statistical issues that may affect its ability to accurately quantify an individual's credit risk (Avery *et al.*, 2000).

In the last decade, there have been moves both to expand and to unify the objectives of consumer credit models (Thomas *et al.*, 2005). Hence, a great number of relevant models have been developed for enhancing banks' ability to effectively evaluate risks associated with loan or credit card applications. Within this framework several methodologies have advanced, including linear discriminant analysis (Saunders, 1977; Long, 1976; Lee, 1985; Hand and Henley, 1997), logit models, probit models, multivariate regressions, usually in a stepwise manner, logistic regression (Myers and Forzy, 1963; Charterjee and Barcum, 1970; Beranek and Taylor, 1976; Long, 1976; Wiginton, 1980; Campbell and Dietrich, 1983; Gardner and Mills, 1989; Lawrence and Arshadi, 1995; Yobas *et al.*, 2000). There are several studies comparing the results derived from different models (Wiginton, 1980; Charitou, Neophytou and Charalambous, 2004; Chandy and Duett, 1990). In contrast to the parametric methods, the non-parametric and other simple methods can be utilized and employed with missing values and multicollinearity among variables. However, they contain several computational demands (Chaterjee

and Barcum, 1970; Breiman et al., 1984; Henley and Hand, 1996). In principle, there is no agreement among the authors on the most appropriate model used, as many studies resulted in contradictory findings and suggestions.

The main purpose of this paper is to evaluate and test the statistical technique of discriminant analysis on credit card customers' data of a Greek commercial bank and examine whether it is possible to create a model evaluating the credibility of prospective credit-card customers. More specifically, this paper attempts to determine whether there exists a relationship between the on-time payments of credit card customers and their demographic characteristics (particular personal and family status).

The remaining contents of this paper are organized as follows. Section I presents the sample used in the present work. Section II describes the research methodology. Section III discusses the empirical results obtained. Section IV provides a summary of the paper.

2. Research Samples

Overall, the sample size in credit scoring studies varies from a few hundred to some tenths of thousands observations (Avery *et al.*, 2004). For example, Duffy (1977) asserts that as few as 300 observations are adequate to develop consumer credit models.

Data-base includes personal, as well as, payment consistency data, over a five year period, for 16,460 customers of the X-BANK, a Greek commercial bank of average size. As Table A1 (Appendix A) reveals, there exist 45 demographic characteristics per customer. However, only a few of them are eligible for inclusion in this analysis, given that most of them are not trustworthy (i.e. annual income) or easily quantifiable (i.e. profession). The omission of some variables may provide a limitation in predicting credit score modeling since credit default may be also driven by the omitted variables (Avery *et al.*, 2000). Based on the most common found in past literature demographic characteristics and after a detailed valuation of the existing data, 14 of these characteristics were selected (Table 1) while the total sample used consists of 1767 customers. Applicants who were rejected were not included in the data. Thus, the sample might be biased as 'good' customers are represented heavily (Hand, 2001). A similar number of variables were, also, used in the earlier study of Smalley and Sturdivant (1973).

The five year sampling period was divided into two equal sub-periods of 2,5 years with the use of a total sample of 1767 customers. The first sub-period aimed at establishing the model using a sample of 829 customers and contained 318 inconsistent paying customers; the second sub-period aimed at validating the model using a sample of 938 customers and contained 360 inconsistent paying customers. In this way, it will be easier to determine the different characteristics of non-payers and 'prompt' payers, which will contribute to the formulation of a desired credit scoring system to rank customer credibility.

3. Research Methodology

The relevant analysis of the sample data includes a comparison of the actual and theoretical values of the bank's customer demographic characteristics, analysis of variance (ANOVA) and the estimation of discriminant analysis' coefficients.

A. Comparison of the actual and theoretical values of the customer characteristics

This methodology includes the following steps:

- The frequencies of appearance of the x customer characteristic (i.e. marital status) are estimated for all non-payers and 'prompt' payers.
- Based on the results of the above calculations, the theoretical frequency of the appearance of the x characteristic is determined for all 'bad' and 'prompt' payers included in the sample.
- The theoretical frequencies are compared with the corresponding actual frequencies of the sample.
- Based on the observed differences, the characteristics of customers closely associated with 'prompt' and 'bad' payers are detected.

The analysis of variance is also used to compare the mean values of the two distributions (theoretical and actual) by analyzing comparisons of distributional variance estimates. Strong differences imply that there exist certain characteristics of the customers that are systematically related to their consistency.

B. Linear Discriminant Analysis

The Linear Discriminant Analysis (LDA) was developed by Fisher (1936) who suggested that the best way to separate two groups is to find the linear combination of explanatory variables which provides the maximum distance between the means of two groups. LDA function for two variables can be defined as a linear combination of discriminating (independent) variables, such that:

$$Y_i = a_1 X_1 + a_2 X_2 + \dots + a_n X_n \quad (1)$$

where Y_i = a variable indicating groups, in this analysis consistent and inconsistent payments, a_1, a_2, \dots, a_n = the discriminant coefficients, and X_1, X_2, \dots, X_n = the explanatory variables.

The advantages of LDA are its simplicity and that it can be easily estimated. The proposed method is based upon the assumption of normality distributed data. However, Reichert et al., (1983) proposed that the non-normality of credit information does not provide a limitation for the empirical utilization of the method. Another argument is that the problem of non-normality can be overcome by using a logit model (Wiginton, 1980), a model that is not selected for examination in this paper.

In this case, the purpose of the LDA is to construct a scheme, based upon the set of the n explanatory variables, that separates observations to appropriate groups and describe the overlaps between the groups (Lee, 1985; Eisenbeis and Avery, 1972). More specifically, the methodology of using the LDA to derive a credit score model contains the following steps. First, we select a sample of previous customers of the X-BANK and classified them as 'good' and 'bad' depending on their financial records over a specified period. Second, some demographic characteristics of the customers are selected. Third, the method of LDA is applied on the data to produce a credit scoring model. Finally, a validation sample is considered and Equation (1) is used to compute each customer's credit score. The credit score is compared to a cut-off point to determine the classification of each consumer as 'good' or 'bad'.

Initially, the correlation coefficients between the 14 selected characteristics of X-BANK customers were estimated. The higher correlation coefficient was 0.34 (revealing high relationship between the variables 'age' and 'number of family members') and the lower one was between -0.13 (revealing negative relationship between the variables 'marital status' and 'years in his/her own house').

4. Research Results

4.1. Comparison of Theoretical and Actual Values

Based on the population frequencies for each characteristic of sample customers (829), Table 3 shows the expected against the observed number of 'bad' payers per characteristic, in the subgroup of 'bad' payers (318). For example, given that 34% of customers are women, the expected number of women 'bad' payers in the sample is 108, against the actual of 136. The fact that the actual percentage is higher than the theoretical proves that there is a tendency for women to be less prompt in their payments.

Table 4 relates the two values by presenting the ratio of actual value to theoretical value. For instance, the third ratio in the ranking (0.91) shows that 'non-prompt' payers who are not married were 91% of the expected number, clearly indicating that this customer category is characterized by high credibility.

The estimations resulting from Table 4 are summarized in Table 5. The latter describes all customer categories (men, women, married, single, home owners, renters, etc.) in terms of their credibility ranking. There are indications on the relationship between the level of customer promptness in payments and their individual characteristics.

Therefore, as a minimum contribution to the analysis, Table 5 provides X-BANK's approving bodies with information on which characteristics they should pay more attention. Also, Table 5 introduces the independent variables that should be examined in the context of developing a certain rating system, through a differential or other form of analysis. In order to evaluate the relationship that exists between actual lack of promptness in payments and theoretically expected 'bad' payers, the following scales have been created:

- Characteristics with ratios in the range 0.9-0.11 are regarded as of average credibility (grey are)
- Characteristics with ratios in the range 0.7- 0.9 are regarded as corresponding to 'prompt' credit card payers.
- Characteristics with values less than 0.7 are regarded as corresponding to very 'prompt' credit card payers.
- Characteristics with value in the range 0.11 to 0.3 are thought to correspond to non 'prompt' credit card owners.
- Characteristics with values over 1.30 reveal extremely 'bad' credit card payers.

The classification in Table 5 can be varied at will, as it is not based on a commonly accepted classification methodology. Yet, regardless of the intervals that will be determined in order to classify customers with different credibility rates, it is essential that these intervals are broad enough to ensure a clear differentiation of the various credibility rankings.

The results from the ANOVA in sample findings show that the theoretical and actual values of the examined characteristics (14x2 characteristics) are not substantially different. Table 6 presents the F-test value.

This result remains after performing two further statistical relevance tests for the two variables (theoretical and actual). Their correlation coefficient is 84.7% and it is statistically significant at the significant level of 1%. Additionally, the regression between the actual and theoretical values has a statistically significant and relatively high R^2 (71.8 %).

4.2. Linear Discriminant Analysis

In order to determine any systematic relationships between the individual characteristics of credit card owners and their rate of response to overdue liabilities, the method of LDA was employed. Theoretically, if the covariance matrices of the underlying populations are unequal, then quadratic discriminant analysis (QDA) should be employed. However, the latter seems to be more sensitive to the model assumptions than LDA and so several authors have concluded the robustness of the LDA over the QDA, including Dillon and Goldstain (1984) and Sharma (1996). The LDA was conducted by using all 14 variables and also by utilizing various combinations of these variables.

4.3. Model with all variables

The discriminant model using the initial sample of 829 customers has a low R^2 (16.15 %), a fact that forejudges a limited capability of independent variables to explain the level of customer credibility. Yet, the discriminant function's coefficients are almost all statistically significant at the significance level of 5%.

For this reason, we have proceeded with testing the model in the context of the validation sample comprising of customers with delays between 2 and 7 months. The validation sample has size 938. Then, the values of the coefficients were applied on the validation sample data, in order to estimate the theoretical values of Y_i per customer.

The results of this validation process are summarized in Table 7 where it is clearly indicated that the model that uses all the selected 14 characteristics as independent variables cannot be used to classify these customers as 'prompt' or 'bad' payers. On the contrary, the results from the use of this model almost equal those of a random distribution: 45% and 55% against 50% and 50%.

In order to strengthen the likelihood for obtaining accurate classification results using the validation sample, the process was repeated by establishing a grey zone with value limits at 0.3 and 0.5 through a fuzzy process. Yet, neither this approach could lead to the formulation of a strong explanatory model (Table 8).

After determining the model in which all sample variables were included and given the fact that this model did not produce the desired results, a stepwise analysis for formulating alternative models was conducted with main criteria the coefficient t-test values, R^2 and the F-test of the corresponding regression lines (Draper and Smith (1981); Jennich (1977a); Jennich (1977b)). The validation process on the 938-customer sample was crosschecked, considered also 73 different combinations of the independent variables, as in the case of the previous model used.

The best results were obtained from a model having ten explanatory variables, like Sex, Marital status, Dependants, Home owner/ or rent, Years living in property, Years in the same job, Place of account delivery, C-Card, X-Bank staff and X-Bank group staff. Using this model 'bad' payer classification is correct in 59% of instances. The corresponding percentage for 'prompt' payers was 54% (Table 9). All the coefficients of the discriminant function are statistically significant at the significance level of 5%.

Therefore, the risk of non-accurate classifications, i.e. the likelihood of 'bad' payers to be accepted as 'prompt' payers and vice versa is high.

Finally, the procedure stated above was repeated for the variables corresponding to the characteristics of 'prompt' payers and very 'bad' payers (see Table 5). The results indicated that these variables could not be utilized to formulate, via discriminant analysis, a model evaluating the credibility of prospective card customers (% of correctly classified as consistent 46% and 51% of correctly classified as inconsistent).

At this point, it is noted that among the reasons for the lack of systematic relationships between customers and paying their debts on time are the following:

a) Most of the data was qualitative with values between 0 and 1. Therefore, the data cannot reflect small changes as in the case of quantitative data.

b) The temporal duration of the sample is likely to be inadequate. The analysis of the data for a longer period of time would probably improve the results.

c) During the past 5 years, stronger competition in the banking sector has led banking institutions to issue credit cards without examining customer data. Moreover, no age group was excluded, resulting in cards being issued to people from 20 to 70 years old.

Given that credit cards were selectively issued during the past 20 years, now credit cards were distributed to all remaining customers, who most probably constitute a random sample, with consequences in the accuracy of related models.

Additionally, note that delays in settling banking liabilities do not necessarily mean that a person is not a 'prompt' payer; delays can be a sign of neglect or belief that no problem is caused by such behavior. It is possible that considering these customers as 'bad' payers is wrong, as delay to pay their debts does not cause damage but is subject to interest rate charges, which constitute additional revenue for banking institutions. In such cases, the bank is the one to decide whether it prefers revenues to liquidity or the opposite. The fact that credit card interest rates are much higher than the interest rate of similar products means that such products are more advantageous and a delay in their payments is to the bank's benefit. If the bank decides that this situation is profitable, it should not seek those customers who will pay on time, but those who delay their payments but eventually pay.

5. Conclusions

The process of credit scoring is very important for banks as they need to segregate 'good' credit-card payers from 'bad' card payers in terms of their creditworthiness. The present work involves the data analysis of credit card owners of a Greek Commercial Bank of average size with the purpose to examine the possibility of the creation of a model evaluating the credibility of prospective credit card customers. After using linear discriminant analysis, it became apparent that establishing such a desired model is not possible based on this method.

The failure of the credit scoring models used in this study to evaluate information of bank customers' demographic characteristics (economic and personal data) raises important statistical issues affecting their prediction accuracy and exhibits their relative potential value and practical limitations in the everyday business life.

Despite the results derived from using linear discriminant analysis, there is no doubt that banks will continue to employ credit scoring based on more sophisticated statistical models as a major tool in predicting credit risk and thus gaining strategic advantages over its competitors. The current global economic crisis enhances the need for an early risk identification system (credit-card scoring model), alerting Commercial Banks against all those prospective customers who may suddenly become default or delinquent.

References

- Avery, R., Bostic, R., Calem, P. and Canner, G. (2000). Credit Scoring: Statistical issues and Evidence from Credit-Bureau Files. *Real Estate Economics*, Vol. 3, 523-547.
- Avery, R., Calem, P. and Canner, G. (2004). Consumer credit scoring: Do situational circumstances matter?, *Journal of Banking and Finance*, Vol. 28, 835-856.
- Beranek, W. and Taylor, W. (1976). Credit scoring models and the cut-off point- A Simplification. *Decision Sciences*, Vol. 7 (July), 394-404.
- Breiman, L., Friedman J.H., Olshen, R.A. and Stone, C.J. (1984). Classification and Regression Trees. Pacific Grove, CA: Wadsworth.
- Campbell, T.S. and Dietrich, J.K. (1983). The Determinants of Default on Insured Conventional Residential Mortgage Loans. *Journal of Finance*, 38, 1569-1581.
- Capon, N. (1982). Credit Scoring Systems: A Critical Analysis. *The Journal of Marketing*, Vol. 46, 82-91.
- Chandy, P.R., Duett, E.E. (1990). "Commercial Papers Rating Models", *Quarterly Journal of Business and Economics*, Vol. 29, pp. 79-101.
- Charitou, A., Neophytou, E. and Charalambous, C. (2004). Predicting Corporate Failure: Empirical Evidence from UK. *European Accounting Review*, Vol. 13, 465-497.
- Charterjee, S. and Barcum, S. (1970). A non-parametric approach to credit screening. *Journal of the American Statistical Association*, Vol. 65 (March), 150-154.

- Dillon, W.R. and Goldstein, M. (1984). "Multivariate Analysis Methods and Applications. Wiley, New York, NY.
- Draper, N. R. and Smith, H. (1981). Applied regression analysis. New York: Wiley.
- Duffy, W. (1977). The Credit Scoring Movement. *Credit* (September), 28-30.
- Eisenbeis, R.A. and Avery, R.B. (1972). Discriminant Analysis and Classification Procedures: Theory and Applications. D. C. Health and Company, Lexington.
- Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, Vol. 7, 179-188.
- Gardner, M.J. and Mills, D.L. (1989). Evaluating the Likelihood of Default on Delinquency Loans. *Financial Management*, Vol.18, 55-63.
- Hand, D.J., (2001). Modeling consumer credit risk. *IMA Journal of Management Mathematics*, Vol.12, 139-155.
- Hand, D.J. and Henley, W.E. (1997). Statistical Classification Methods in Consumer Credit Scoring. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Vol. 160, 523-541.
- Henley, W.E. and Hand, D.J. (1996). A k-nearest Neighbour Classifier for Assessing Consumer Credit Risk. *Statistician*, Vol. 45, 77-95.
- Jennrich, R. I. (1977a). Stepwise regression. In K. Enslein, A. Ralston, & H. S. Wilf (Eds.). Statistical methods for digital computers, New York: Wiley, 58-75.
- Jennrich, R. I. (1977b). Stepwise discriminant analysis. In K. Enslein, A. Ralston, & H. S. Wilf (Eds.). Statistical methods for digital computers, New York: Wiley, 76-96.
- Lawrence, E. and Arshadi, N. (1995). A Multinomial Logit Analysis of Problem Loan Resolution Choices in Banking. *Journal of Money, Credit and Banking*, 27, 202-216.
- Lee, C.F. (1985). Financial analysis and planning: Theory and Application. Addison-Wesley Publishing Company, Reading Massachusetts.
- Long, M. (1976). Credit scoring system selection. *Journal of Financial and Quantitative Analysis*, 11 (June), 313-328.
- Mester, L. (1997). What's the Point of Credit Scoring?. *Business Review*, Federal Reserve Bank of Philadelphia, September/October, 3-16.
- Myers, J. and Forzy, E. (1963). The Development of Numerical Credit Evaluation Systems. *Journal of the American Statistical Association*, 58 (September), 799-806.
- Nevin, J. and Churchil, G. (1979). The Equal Credit Opportunity Act: An Evaluation. *Journal of Marketing*, 43, 95-104.
- Reichert, A.K., Cho, C.C. and Wagner, G.M. (1983). An examination of Conceptual Issues Involved in Developing Credit-scoring Models. *Journal of Business and Economic Statistics*, Vol. 1, 101-114.
- Saunders, A. (1977). Financial Institutions Management: A Modern Perspective. Boston: Irwin, Boston.
- Sharma, S. (1996). Applied Multivariate Techniques. Wiley, New York, NY.
- Smalley, O. and Sturdivant, F. (1973). The Credit Merchant: A History of Spiegel Inc. IL: Southern Illinois University Press.
- Thomas, L.C., Oliver, R.W. and Hand D.J. (2005). A survey of the issues in consumer credit modelling research. *Journal of the Operational Society*, Vol. 56, 1006-1015.
- Wiginton, J. C. (1980). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behaviour. *Journal of Financial and Quantitative Analysis*, Vol. 15, 757-770.
- Yobas, M.B., Crook, J.N. and Ross, P. (2000). Credit Scoring Using Neural and Evolutionary Techniques. IMA, Journal of Mathematical Applications in Business Industry, Vol. 11, 111-125.

Table 1. Demographic Characteristics of X-BANK' s Customers

1	Gender	C*	Male, Female
2	Marital status	C	Married, Single, Other
3	Dependants	N	
4	Date of birth	C	
5	Home owner/ or rent	C	
6	Years living in own property	N	
7	Profession code	N	
8	Owens a car	C	YES, NO
9	Holds an insurance	C	YES, NO
10	Bills sent at home	C	YES, NO
11	Automatic payment of bills	C	YES, NO
12	C Card	C	YES, NO
13	XBANK staff	C	YES, NO
14	Bank Group staff	C	YES, NO

* C= Character, N= Numeric

Table 2. Personal Characteristics to be Analysed

Credibility category	Payment overdue for	Number of customers in the sample
1	Up to 1 month	10.278
2	1-2 months	2.837
3	2-3 months	1.234
4	3-4 months	754
5	4-5 months	568
6	5-6 months	470
7	Over 6 months	319

Table 3. Theoretical Number of Inconsistent Customers

Total number of customers = 829			
Number of inconsistent customers in the sample = 318			
	Population Frequency*	Theoretical number of inconsistent customers	Actual number of inconsistent customers
P(Sex: woman =0)	34%	108	136
P(Sex: man= 1)	66%	210	182
P(Marital status: single=0)	50%	159	144
P(Marital status: married=1)	50%	159	174
P(Family: protective members=0)	67%	213	BO
P(Family: no protective members=1)	33%	105	138
P(Age<31)	32%	102	86
P(Age 31-45)	47%	149	170
P(Age 46-60)	17%	54	52
P(Age >60)	4%	13	10
P(house status: rented = 0)	40%	127	161
P(house status: owned = 1)	60%	191	157
P(Car ownership: No=0)	56%	178	152
P(Car ownership: Yes=1)	44%	140	166
P(XBANK group staff :Yes = 1)	6%	19	36
P(Years at the same home 0-9)	5%	16	195
P(Years at the same home 10-19)	13%	41	69
P(Years at the same home 20-29)	21%	67	45
P(Years at the same home >30)	61%	194	9
P(Years at the same job 0 - 9)	83%	264	242
P(Years at the same job 10-19)	12%	38	55
P(Insurance contract: No=0)	55%	175	171
P(Insurance contract: Yes=1)	45%	143	147
P(Place of account delivery : work=0)	20%	64	89
P(Place of account delivery : home=1)	80%	254	229
P(Bank payment: No= 0)	88%	280	300
P(Bank payment: Yes=1)	12%	38	18
P(C Card : No = 0)	75%	239	240
P(C Card : Yes=1)	25%	80	78
P(XBANK staff :No =0)	98%	312	317
P(XBANK staff :Yes=1)	2%	6	1
P(XBANK group staff :No = 0)	94%	299	282

* The frequencies derived by using the demographic characteristics of all 16,460 customers.

Table 4. Actual to Theoretical Values

	Theoretical value	Actual value	Actual / Theoretical value
P(Sex: woman =0)	108	136	1.26
P(Sex: man= 1)	210	182	0.87
P(Marital status: single=0)	159	144	0.91
P(Marital status: married=1)	159	174	1.09
P(Family: protective members=0)	213	180	0.84
P(Family: no protective members=1)	105	138	1.32
P(Age<31)	102	86	0.85
P(Age 31-45)	149	170	1.14
P(Age 46-60)	54	52	0.96
P(Age >60)	13	10	0.79
P(house status: rented = O)	127	161	1.27
P(house status: owned = 1)	191	157	0.82
P(Car ownership: No=0)	16	195	12.26
P(Car ownership: Yes=1)	41	69	1.67
P(XBANK group staff :Yes = 1)	67	45	0.67
P(Years at the same home 0-9)	194	9	0.05
P(Years at the same home 10-19)	264	242	0.92
P(Years at the same home 20-29)	38	55	0.92
P(Years at the same home >30)	16	21	1.32
P(Years at the same job 0 - 9)	178	152	0.85
P(Years at the same job 10-19)	140	166	1.19
P(Insurance contract: No=0)	175	171	0.98
P(Insurance contract: Yes=1)	143	147	1.03
P(Place of account delivery : work=0)	64	89	1.40
P(Place of account delivery : home=1)	254	229	0.90
P(Bank payment: No= 0)	280	300	1.07
P(Bank payment: Yes=1)	38	18	0.47
P(C Card : No = 0)	239	240	1.01
P(C Card : Yes=1)	80	78	0.98
P(XBANK staff :No =0)	312	317	1.02
P(XBANK staff :Yes=1)	6	1	0.16
P(XBANK group staff :No = 0)	299	282	0.94
P(Sex: woman =0)	19	36	1.89

Table 5. Customers' Classification According to their Consistency

	Actual / Theoretical value	
P(Sex: woman =0)	1.26	Inconsistent
P(Sex: man= 1)	0.87	Consistent
P(Marital status: single=0)	0.91	Average
P(Marital status: married=1)	1.09	Average
P(Family: protective members=0)	0.84	Consistent
P(Family: no protective members=1)	1.32	Highly inconsistent
P(Age<31)	0.85	Consistent
P(Age 31-45)	1.14	Inconsistent
P(Age 46-60)	0.96	Average
P(Age >60)	0.79	Consistent
P(house status: rented = 0)	1.27	Inconsistent
P(house status: owned = 1)	0.82	Consistent
P(Car ownership: No=0)	12.26	Highly inconsistent
P(Car ownership: Yes=1)	1.67	Highly Consistent
P(XBANK group staff :Yes = 1)	0.67	Highly consistent
P(Years at the same home 0-9)	0.05	Highly consistent
P(Years at the same home 10-19)	0.92	Average
P(Years at the same home 20-29)	0.92	Average
P(Years at the same home >30)	1.32	Highly inconsistent
P(Years at the same job 0 - 9)	0.85	Consistent
P(Years at the same job 10-19)	1.19	Inconsistent
P(Insurance contract: No=0)	0.98	Average
P(Insurance contract: Yes=1)	1.03	Average
P(Place of account delivery : work=0)	1.40	Highly Inconsistent
P(Place of account delivery : home=1)	0.90	Average
P(Bank payment: No= 0)	1.07	Average
P(Bank payment: Yes=1)	0.47	Highly consistent
P(C Card : No = 0)	1.01	Average
P(C Card : Yes=1)	0.98	Average
P(XBANK staff :No =0)	1.02	Average
P(XBANK staff :Yes=1)	0.16	Highly consistent
P(XBANK group staff :No = 0)	0.94	Average
P(Sex: woman =0)	1.89	Inconsistent

Table 6. ANOVA (Theoretical to Actual values of the sample)

F-statistic	F-critical value
0.0008	3.99*

* The degrees of freedom are 1 (between groups) and 64 (within groups). The significance level is 5%.

Table 7. Customers' Classification using as Cut-off point 0.388*

	Consistent	Inconsistent	Total
Customers in the validation sample	578	360	938
Correctly classified as consistent (using the discriminant model with the initial sample)	(62%) 287	(38%) 199	(100%) 486
Correctly classified as inconsistent (using the discriminant model with the initial sample)	50%	55%	
Total customer correctly classified			
% of correctly classified as consistent			
% of correctly classified as inconsistent			

* Values less than 0.388 correspond to 'prompt' payers, while the opposite is indicated by values over 0.388

Table 8. Customers' Classification using the Grey zone 0.3 - 0.5

	Consistent	Inconsistent	Total
Customers in the validation sample	578	360	938
Correctly classified as consistent (using the discriminant model with the initial sample)	(62%) 150	(38%) 105	(100%) 255
Correctly classified as inconsistent (using the discriminant model with the initial sample)	26%	29%	
Total customer correctly classified			
% of correctly classified as consistent			
% of correctly classified as inconsistent			

Table 9. Customers' Classification using a Cut-off point 0.398

	Consistent	Inconsistent	Total
Customers in the validation sample	578	360	938
Correctly classified as consistent (using the discriminant model with the initial sample)	(62%) 340	(38%) 194	(100%) 534
Correctly classified as inconsistent (using the discriminant model with the initial sample)	59%	54%	
Total customer correctly classified			
% of correctly classified as consistent			
% of correctly classified as inconsistent			

Appendix A

Table A1. Demographic Characteristics of X-BANK' s Customers

1	Application origin	C	
2	Gender	C	Male, Female
3	Nationality	C	
4	Marital status	C	Married, Single
5	Dependants	N	
6	Date of birth	C	
7	Postal code	N	
8	Home owner/ or rent	C	
9	Years living in property	N	
10	Profession code	N	
11	Years working for employer	N	
12	Work postal code	N	
13	Annual personal income	N	
14	Family income	N	
15	Other income	C	
16	Other sources of income	C	
17	He/she has an account in other banks	C	YES, NO
18	XBANK customer	C	YES, NO
19	He/she has a savings account	C	YES, NO
20	He/she has a current account	C	YES, NO
21	He/she has mutual funds	C	YES, NO
22	He/she has a housing loan	C	YES, NO
23	He/she has a consumer loan	C	YES, NO
24	Other banking products	C	YES, NO
25	Other products (description)	C	
26	He/she has a BANK 1 Card	C	YES, NO
27	He/she has a BANK 2 Card	C	YES, NO
28	He/she has a BANK 3 Card	C	YES, NO
29	He/she has a BANK 4 Card	C	YES, NO
30	He/she has a BANK 5 Card	C	YES, NO
31	He/she has a BANK 6 Card	C	YES, NO
32	He/she has a BANK 7 Card	C	YES, NO
33	He/she has a BANK 8 Card	C	YES, NO
34	He/she has a BANK 9 Card	C	YES, NO
35	He/she has a BANK 10 Card	C	YES, NO
36	He/she has a car	C	YES, NO
37	Year the car was bought	N	
38	He/she has an insurance	C	YES, NO
39	Bills sent at home	C	YES, NO
40	Automatic payment of bills	C	YES, NO
41	Balance	N	
42	C Card	C	YES, NO
43	Card number	N	
44	XBANK staff	C	YES, NO
45	Bank Group staff	C	YES, NO
C= character, N=numeric			