12-1-2016 12:00 AM

Applications of Credit Scoring Models

Mimi Mei Ling Chong
The University of Western Ontario

Supervisor
Dr. Matt Davison
The University of Western Ontario

Graduate Program in Statistics and Actuarial Sciences
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Abstract

The application of credit scoring on consumer lending is an automated, objective and consistent tool which helps lenders to provide quick loan decisions. In order to apply for a loan, applicants must provide their attributes by filling out an application form. Certain attributes are then selected as inputs to a credit scoring model which generates a credit score. The magnitude of this credit score is proved to be related to the credit quality of the loan applicant. As such, it is used to determine whether the loan will be granted, and also the amount of interest being charged. Currently, little effort has been devoted to verifying the correctness of the reported attributes provided by prospective borrowers. Moreover, with a long history of use of the same credit scoring model, borrowers will learn about the characteristics being used by the lender to make loan decisions, and may be motivated to lie about their attributes in order to increase their chances of loan approval. This thesis examines the effect on consumer lending if some borrowers strategically falsify their attributes on the application form. Under normal circumstances, analysts believe that using a larger dataset to develop credit scoring models will increase model accuracy. We will show that if some borrowers respond dishonestly to some questions on the application form, using higher dimensional data to build models will increase the associated accumulated error, and may result in having a more complex model but with low predictive power compared against using a dataset with lower dimensions. Furthermore, we will show that it is profitable for lenders to spend extra cost to directly eliminate lies in the dataset. In particular, we will examine the optimal amount of effort that lenders should spend on identifying liars in order to equilibrate between risk and return. However, we will also show that it is still possible for fraudulent loan applicants to eventually adjust their lies to escape from credit checks and get loans. Indeed the business of consumer lending may usefully be modeled as a game performed between the lender and the borrower. We will explore the cost to make a clever lie on the attributes and the cost to verify the correctness of the reported data towards the interaction between the lender and bad liars. The impact of having liars in the business not only affects the profitability of lenders, but also lowers the utility of those borrowers who always repay their loans and the utility of the economy as a whole. The proposed issues will be studied using discriminant analysis on simulated data, and then further assume the characteristics of borrowers follow half triangular distribution to present theoretical results. This research shows the importance of enriching data before making loan decisions. It can help lending financial institutions to reduce risk and maximize profit, and it also shows that it is feasible for customers to lie intelligently so as to evade credit checks and get loans.

Keywords: Credit Scoring, Discriminant Analysis, Fraud Detection, Utility, Default

Co-Authorship Statement

I hereby declare that this thesis incorporates materials that are co-authored with Dr. Matt Davison and Dr. Cristián Bravo. Dr. Davison is a Professor in the Department of Applied Mathematics and Statistical & Actuarial Sciences at Western University and he is my supervisor. Dr. Bravo is a Lecturer (Assistant Professor) in Business Analytics within the Department of Decision Analytics and Risk, Southampton Business School at the University of Southampton and acted as my supervisor in this thesis. Details of the co-authored papers are as below:

The content of Chapter 2 appeared as a published conference paper, co-authored with my supervisor, Dr. Davison. "Larger Datasets Lead to More Inaccurate Credit Scoring", *The International Statistical Institute*, (Section CPS009), 3429-3434.

The content of Chapter 4 was a published paper, co-authored with Dr. Davison and Dr. Bravo, "How Much Effort Should Be Used to Detect Fraudulent Applications When Engaged in Classifier-Based Lending?", *Intelligent Data Analysis*, vol. 19, no. s1, pp. S87-S101.

I am the primary author of all the research work presented in this thesis. I was in charge of model implementation, data simulation, results and analysis, literature review and completing the first draft of the manuscripts. Dr. Davison and Dr. Bravo provided valuable insights and innovations on the modeling framework formulation. I certify that this dissertation is fully a product of my own work.

Acknowledgements

I would like to sincerely thank you all the people who helped in the various stages of this research. First and foremost, I would like to gratefully acknowledge my supervisor Dr. Matt Davison for his guidance and indispensable support in completing this research. I greatly admire him for his intelligence and hardworking attitude, which motivates me to take up more challenges in my life. His advice on both my research and career were valuable. I greatly appreciate him for his patience and accessibility.

Sincere thanks to Lynette D'souza and Regina Camaya of CitiFinancial Canada for providing me an opportunity to work with them and to learn more about the retail lending business in the real world.

I would like to extend my warmest thanks to Dr. Cristian Bravo for his valuable suggestions and contributions to our paper. I would also like to thank Dr. Rogemar Mamon for his directions on my teaching assistant role.

A very special thanks goes to Dr. Kim-Fai Hung who used to work at the Hong Kong Polytechnic University until he passed away in 2014, without his motivation and encouragement I would not have considered to pursue a PhD.

Finally, my deepest gratitude goes to my parents and my four older siblings for their perpetual love and constant support throughout my life. I would like to particularly extend my deepest appreciation to my eldest sister Alice for her guidance and unconditional support throughout my growth. Without all your love, encouragement and understanding, it would never be possible for me to complete this PhD study.

Contents

Abstract **ii**

Co-Authorship Statement **iii**

Acknowledgements **iv**

List of Figures **viii**

List of Tables **xvi**

List of Appendices **xx**

1 Introduction **1**

1.1 Background . . . 1

1.1.1 Consumer Credit . . . 1

1.1.2 What is Credit Scoring . . . 5

1.1.3 Benefits of Credit Scoring . . . 7

1.1.4 Other Applications of Credit Scoring . . . 9

1.2 Research Objectives . . . 10

1.3 Structure of the Thesis . . . 11

2 Methods used in Credit Scoring **16**

2.1 Discriminant Analysis . . . 17

2.1.1 Univariate Normal . . . 19

Case I : $\sigma_{\mathbf{G}}^{\mathbf{2}} = \sigma_{\mathbf{B}}^{\mathbf{2}} = \sigma^{\mathbf{2}}$ . . . 20

Case II : $\sigma_{\mathbf{B}}^{\mathbf{2}} - \sigma_{\mathbf{G}}^{\mathbf{2}} > \mathbf{0}$ . . . 22

Case III : $\sigma_{\mathbf{B}}^{\mathbf{2}} - \sigma_{\mathbf{G}}^{\mathbf{2}} < \mathbf{0}$ . . . 26

2.1.2 Bivariate Normal . . . 29

2.2 Logistic Regression . . . 37

2.3 Conclusions . . . 42

3 More Attributes May Lead to More Inaccurate Credit Scoring **45**
3.1 Introduction . . . 45
3.2 Method and Model . . . 46
3.3 Simulation of Data . . . 46
3.4 Results and Analysis . . . 47
3.5 Conclusions . . . 51

4 How Much Effort Should Be Spent to Detect Fraudulent Applications When Engaged in Classifier-Based Lending? **54**
4.1 Introduction . . . 54
4.2 Method and Model . . . 57
4.3 Simulation of Data . . . 59
4.4 Results and Analysis: For the Banks (Single Loan Amount) . . . 59
4.4.1 Case I: $A = 3$, $k = \$390K$ . . . 61
4.4.2 Case II: $A = 3$, $k = \$1080K$ . . . 63
4.4.3 Case III: $A = 4$, $k = \$520K$ . . . 64
4.4.4 Case IV: $A \in \{1, 1.25, 1.5, ..., 6\}$, Unit Cost = \$130K . . . 66
4.4.5 Case V: $A \in \{1, 1.25, 1.5, ..., 5\}$, Unit Cost = \$260K . . . 68
4.5 Results and Analysis: For the Banks (Attribute Dependent Loan Amount) . . . 70
4.5.1 Case I: A = 3, k = \$390K . . . 71
4.5.2 Case III: A = 4, k = \$520K . . . 73
4.6 For the borrowers: . . . 74
4.7 Conclusions . . . 75

5 Cost Effective Game of Banks and Dishonest Borrowers **82**
5.1 Introduction . . . 82
5.2 Method and Model . . . 85
5.3 Data Description . . . 86
5.3.1 Overview of Triangular Distribution . . . 88
5.3.2 Credit scores of good borrowers . . . 90
5.3.3 Credit scores of bad borrowers . . . 90
Bad truth tellers . . . 91
Bad liars . . . 92
5.4 Modeling Borrower and Lender Objective Functions . . . 93
5.4.1 The cutoff score . . . 93
5.4.2 The number of loans granted . . . 94

5.4.3 The utility function of different parties 95
For the banks 96
For the borrowers 97
5.5 Parameters and Analysis 99
5.5.1 Fixed Parameters 99
5.5.2 Modeling Cost of Lying and Cost of Checking 100
5.6 Results and Analysis 103
5.6.1 Scenario 1: Not Economical to Lie 103
5.6.2 Scenario 2a: Economical to Lie But Not Economical to Check for Lies . 107
5.6.3 Scenario 2b: Always Economical to Lie and Check for Lies 111
5.6.4 Scenario 2c: Economical to Lie But Check for Lies Periodically 115
5.7 Conclusions 119

6 Conclusions **125**
6.1 Summary 125
6.2 Business Insights from this work 127
6.2.1 Further research directions 127

A Data Simulation **130**

B Lending Club **133**

C Glossary of Terms **135**

Curriculum Vitae **138**

List of Figures

1.1 This plot shows that Canada's total household debt, mortgage debt and consumer debt have been increasing from 1980 to 2015. 4

1.2 This plot shows the ratio of household credit market debt (mortgages plus consumer credit) to disposable income in Canada from 1990 to 2015. The ratio of household debt to disposable income rose from 85.2% in 1990 to 165.4% in 2015. This means that as of 2015, Canada households owed more than $1.65 in debt for every dollar of annual disposable income. 5

2.1 Data were simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = P_B = 0.5$. The solid line showing the probability density function (PDF) of X_G and dotted line showing the PDF of X_B. The vertical line shows the cutoff value $\tau = 6.40$ obtained by putting $L = 10$, $Q = 2$, $\mu_G = 8$, $\mu_B = 4$ and $\sigma_G = \sigma_B = \sigma = 1$ into Eq. (2.9). The shaded region shows all the borrowers with characteristics $X_1 > \tau$, and will be classified as good payers by set condition (2.8). 22

2.2 Data were simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,2)$ with $P_G = P_B = 0.5$. The solid line showed the probability distribution function (PDF) of X_G, while the dotted line showed the PDF of X_B. The cutoff values were obtained by putting $L = 10$, $Q = 2$, $\mu_G = 8$, $\mu_B = 4$, $\sigma_G = 1$ and $\sigma_B = 2$ into Eq. (2.15) and (2.16). Note that when $\sigma_B^2 > \sigma_G^2$, there were two cutoff points correspond to two vertical cutoff lines with values $a = 7.17$ and $b = 11.49$. The shaded region showed all the borrowers with characteristic values between a and b and was classified as good payers by set condition (2.14). 24

2.3 Data were simulated from $X_G \sim N(8,2)$ and $X_B \sim N(4,1)$ with $P_G = P_B = 0.5$. The solid line displayed the probability distribution function (PDF) of X_G, while the dotted line displayed the PDF of X_B. The cutoff values were obtained by putting $L = 10$, $Q = 2$, $\mu_G = 8$, $\mu_B = 4$, $\sigma_G = 2$ and $\sigma_B = 1$ into Eq. (2.27) and (2.28). Note that when $\sigma_G^2 > \sigma_B^2$, there were two cutoff points with values $a = -0.97$ and $b = 6.31$. The shaded region showed the borrowers with characteristics that will be classified as good payers by set condition (2.26). . . 27

2.4 Data was simulated using $X_{1G} \sim N(8,1), X_{1B} \sim N(4,1), X_{2G} \sim N(8,1), X_{2B} \sim N(4,1), p_B = p_G = 0.5, L = 10,\ Q = 2$ and $\rho = 0$. The solid contour on the top right corner of the plot represents borrowers with characteristics X_{1G} and X_{2G}, while the solid contour on the lower left corner represents borrowers with characteristics X_{1B} and X_{2B}. The dotted contour on the top left corner represents borrowers with characteristics X_{1B} and X_{2G}, and the dotted contour on the lower right corner represents borrowers with characteristics X_{1G} and X_{2B}. The diagonal line is the cutoff line plotted using Eq. 2.41 and the shaded region shows borrowers having characteristics that will be classified as good payers by the set condition 2.39. 32

2.5 An example of logistic function. 40

3.1 Data was simulated using the method described in Section 3.3 and the parameters in Table 3.1. Constant amount of lies was added to X_{2B} according to equation (3.2) to generate X_{2Bnew}. The plot of histogram and Q-Q plot provide evidence that it is possible to misinterpret X_{2Bnew} as normally distributed. . . . 48

3.2 The gray contour on the left displays the characteristic values of the bad payers, while the black contour on the right displays the characteristic values of the good payers. The solid line is the cutoff line plotted using equation (3.4), while the dotted line is the cutoff line plotted using equation (3.5). 52

4.1 Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = 0.8$, $P_B = 0.5$ and $n = 500$ borrowers, η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 3$, $P_N = \frac{1}{2}$, $L = \$7.2K$ and $k = \$390$K to classify borrowers. We used Eq. (4.5) to calculate the revenue for each η. We computed the average revenue of 100 datasets and reran the program 100 times to obtain the average mean values of revenue, together with the 95% confidence interval of the mean values displayed in the plot as the dashed lines. The lower right corner of the plot showed a zoom-in subplot to improve the visualization of the plot. 62

4.2 Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = 0.8, P_B = 0.2$ and $n = 500$ borrowers,η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 3$, $L = \$7.2\text{K}$, $k = \$390\text{K}$ and $P_N = \frac{1}{2}$ to classify borrowers. We used Eqs. (4.5), (4.6) and (4.7) to calculate the increase in revenue and profit for each η. We reran the program 100 times to obtain the average mean values of increase in revenue and profit as stated in the solid and open square lines respectively. The grey dashed lines displayed the 95% confidence interval of the mean values. 64

4.3 Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = 0.8, P_B = 0.5$ and $n = 500$ borrowers, η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 3$, $L = \$7.2\text{K}$, $k = \$1080\text{K}$ and $P_N = \frac{1}{2}$ to classify borrowers. We used Eqs. (4.5), (4.6) and (4.7) to calculate the increase in revenue and profit for each η. We reran the program 100 times to obtain the average mean values of increase in revenue (solid line) and profit (open square line), together with the 95% confidence interval of the mean values, displayed as the dashed lines. 65

4.4 Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = 0.8, P_B = 0.2$ and $n = 500$ borrowers, η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 4$, $k = \$520\text{K}$, $L = \$7.2\text{K}$ and $P_N = \frac{1}{2}$ to classify borrowers. We used Eqs. (4.5), (4.6) and (4.7) to calculate the increase in revenue and profit for each η. We reran the program 100 times to obtain the average mean values of increase in revenue and profit, together with the 95% confidence interval of the mean values, displayed as the dashed lines. 67

4.5 Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = 0.8, P_B = 0.2$ and $n = 500$ borrowers, η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 4.5$, $k = \$520\text{K}$, $L = \$7.2\text{K}$ and $P_N = \frac{1}{2}$ to classify borrowers. We used Eqs. (4.5), (4.6) and (4.7) to calculate the increase in revenue and profit for each η. We reran the program 100 times to obtain the average mean values of increase in revenue and profit, together with the 95% confidence interval of the mean values, displayed as the dashed lines. 68

4.6 The plot of η_{opt} versus A from Table 4.2. 69

4.7 Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = 0.8, P_B = 0.2$ and $n = 500$ borrowers, η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 3$, $P_N = \frac{1}{2}$ and $k = \$390\text{K}$ to classify borrowers. We used Eq. (4.5) to calculate the revenue for each η. We computed the average revenue of 100 datasets and reran the program 100 times to obtain the average mean values of revenue, together with the 95% confidence interval of the mean values displayed in the plot as the dashed lines. The lower right corner of the plot showed a zoom-in subplot to improve the visualization of the plot. 72

4.8 Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = 0.8, P_B = 0.2$ and $n = 500$ borrowers, η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 3$, $k = \$390\text{K}$ and $P_N = \frac{1}{2}$ to classify borrowers. We used Eqs. (4.5), (4.6) and (4.7) to calculate the increase in revenue and profit for each η. We reran the program 100 times to obtain the average mean values of increase in revenue and profit as stated in the solid and open square lines respectively. The grey dashed lines displayed the 95% confidence interval of the mean values. 73

4.9 Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = P_B = 0.5$ and $n = 500$ borrowers, η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 4$, $k = \$520\text{MM}$ and $P_N = \frac{1}{2}$ to classify borrowers. We used Eqs. (4.5), (4.6) and (4.7) to calculate the increase in revenue and profit for each η. We reran the program 100 times to obtain the average mean values of increase in revenue and profit, together with the 95% confidence interval of the mean values, displayed as the dashed lines. 74

4.10 Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = P_B = 0.5$ and $n = 500$ borrowers. We varied A from 0 to 12 with step size 0.04. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$ and $P_N = \frac{1}{2}$ to generate the table shown in Table 4.5. To compare the effect of η, we set η equals to 0, 0.2 and 0.4. For each value of η, we computed the average odds ratio of 100 datasets. The above plot showed that bad borrowers should lie in the values of A where odds ratio is greater than 1, in order to increase their chance of having loan approval. Furthermore, it showed that increasing the value of η will require a higher value of A to maintain the same level of odds ratio. 76

4.11 Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = P_B = 0.5$ and $n = 500$ borrowers. We varied A from 0 to 12 with step size 0.04. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $\eta = 0.2$ and $P_N = \frac{1}{2}$ to generate the table shown in Table 3. We reran the program for 100 times to obtain the mean average of odds ratios. We found the confidence interval estimates of the average odds ratio, as shown as the dashed lines in the above plot. 77

5.1 The dataset in Lending Club website contains 236K borrowers' records and reported FICO scores as two numbers: FICO low and FICO high. FICO scores of borrowers have been calculated as the average of FICO low and FICO high. This histogram exhibit FICO scores with a positively skewed normal trend. . . 88

5.2 Approximate FICO trend of good and bad payers. 89

5.3 The probability density function of a asymmetric triangular distribution with support [c,d], mode m and height $\frac{2}{d-c}$. 90

5.4 This plot presents the probability density function (PDF) of the credit scores of different borrowers. The solid line displays the PDF of the scores of good borrowers, which follows left triangular distribution and has values varies from $2a$ to $4a$. The dash line shows the PDF of bad truth tellers, which follows right triangular distribution and has value varies from a to $3a$. The solid line with circle pattern shows the PDF of the credit scores of bad liars. It follows right triangular distribution. A constant amount of A represents the amount of lies that bad liars altered their attributes and resulted in an increase on their score. A parameter η represents the amount of effort spend on eliminating lies. The mix effect of A and η causes the credit scores of bad liars to vary from $a + A(1-\eta)$ to $3a + A(1-\eta)$. 91

5.5 The plot above shows that different combinations of h and k results in four different scenarios, each represents a different interaction between the bank and bad liars. We varied h from \$0 to \$3 and k from \$0 to \$70 each with step size 0.1. The top dark gray region represents the scenario where it is not economical to lie. The light gray region represents the scenario where it is economical to lie but not to check for lies. The black and white regions both represents the scenario where it is economical to lie and to check for lies, with the black region representing lying and checking should be done periodically, and white region represents lying and checking should always be carried out. . 102

5.6 This plot presents the probability density function (PDF) of the scores of different borrowers. We set a to 200, A to 0 and η to 0 in Figure 5.3 and added the cut off line $\gamma = 491$. The dash line overlapping the solid line with circle patterns have scores ranging from 200 to 600 represents the PDF of bad honest borrowers and bad liars respectively. The solid line with scores ranging from 400 to 800 represents the PDF of good borrowers. The vertical line with $x = 491$ showed the cutoff value. The gray shaded region showed the applicants with their score greater than the cutoff value and will be granted loans. 106

5.7 The above plot showed the utility of bad liars (in millions) versus different values of A. We applied the parameters in Table 5.1, set $\eta = 0$ and $\gamma = 491$, arbitrarily assigned h to \$3, varied A from 0 to 200 with stepsize 0.01, and for each values of A calculated U_{Blied}. This plot showed that maximum utility of bad liars occur at A equals 0, stating that bad liars should not any generate lie in order to maintain at their maximum utility level. 107

5.8 The above plot showed the utility of bad liars (in millions) versus different values of A. We applied the parameters in Table 5.1, set $\eta = 0$ and $\gamma = 491$, arbitrarily assigned h to \$1, varied A from 0 to 200 with step size 0.01, and for each values of A calculated U_{Blied}. This plot showed that maximum utility of bad liars occur at A equals 200 , stating that bad liars should lie at their maximum to obtain at their maximum utility level. 110

5.9 This plot presents the probability density function (PDF) of the attributes of different borrowers. We set a to 200, A to 200 and η to 0 in Figure 5.3 and added the cut off line $\gamma = 491$. The dash line with attribute values ranging from 200 to 600 represents the PDF of bad honest borrowers. The solid line with circle patterns with attribute values ranging from 400 to 800 represents the PDF of bad liars, while the solid line with attribute values ranging from 400 to 800 represents the PDF of good borrowers. The gray shaded region showed the applicants with their attribute value greater than the cutoff value and will be granted loans. 110

5.10 The cutoff value has been recalculated by substituting $A = 200$, $\eta = 0$ into Eq. (5.10) and resulted with $\gamma = 536$. This plot presents the probability density function (PDF) of the attributes of different borrowers. We set a to 200, A to 200 and η to 0 in Figure 5.3 and added the cut off line $\gamma = 536$. The dash line with attribute values ranging from 200 to 600 represents the PDF of bad honest borrowers. The solid line with circle patterns with attribute values ranging from 400 to 800 represents the PDF of bad liars, while the solid line with attribute values ranging from 400 to 800 represents the attributes of good borrowers. The gray shaded region showed the applicants with their attribute value greater than the cutoff value and will be granted loans. 111

5.11 The above plot showed the utility of the bank (in millions) versus different values of η. We applied the parameters in Table 5.1, set $A = 200$, assigned k to \$70, varied η from 0 to 1 with step size 0.001, and for each values of η calculated U_{BK}. This plot showed that maximum utility of the bank occur at η equals 0 , stating that banks should not spend money to check for lies in applicants' attribute. 112

5.12 The above plot showed the utility of the banks (in millions) versus different values of η. We applied the parameters in Table 5.1, set $A = 200$, $k = \$50$, and varied η from 0 to 1 with step size 0.001. For each values of η we calculated the corresponding U_{BK}. The dash line showed that maximum utility of the banks occurs at $\eta = 0.412$ stating that banks should put 0.412 amount of effort to check for lies in applicants' attribute. 114

5.13 This plot shows the PDF of different borrowers in Step 4. The cutoff value has been calculated by substituting $A = 1$, $\eta = 0.412$ into Eq. (5.10) and resulted with $\gamma = 2.59$. This plot presents the probability density function (PDF) of the scores of different borrowers. We set a to 1, A to 1 and η to 0.412 in Figure 5.4 and added the cut off line $\gamma = 2.59$. The dash line with scores ranging from 1 to 3 represents the PDF of bad honest borrowers. The solid line with circle patterns with scores ranging from 1.59 to 3.59 represents the PDF of bad liars, while the solid line with scores ranging from 2 to 4 represents the scores of good borrowers. The gray shaded region showed the applicants with their scores greater than the cutoff value and will be granted loans. 115

5.14 This plot showed the utility of bad liars (in millions) versus different values of A in Step 4 of Scenario 2b. We applied the parameters in Table 5.1, set $\eta = 0.412$ and $\gamma = 518$, arbitrarily assigned h to \$1 and k to \$50, and varied A from 0 to 200 with step size 0.01. For each values of A we calculated U_{Blied}. This plot showed that maximum utility of bad liars occurred at A equals 1 , stating that bad liars should keep adding maximum amount of lies onto their actual score in order to obtain at a utility level of 4.99MM. 116
5.15 The above plot showed the utility of the banks (in millions) versus different values of η. We applied the parameters in Table 5.1, set $A = 200$, $k = \$20$, and varied η from 0 to 1 with step size 0.001. For each values of η we calculated the corresponding U_{BK}. This plot suggested banks to set η to 1 to put full effort to check for lies in applicants' attribute. The maximum amount of utility that banks can obtained is 193MM. 118
5.16 This plot showed the utility of bad liars (in millions) versus different values of A in Step 5 of Scenario 2c. We applied the parameters in Table 5.1, set $\eta = 1$ and $\gamma = 491$, assigned h to \$1 and k to \$20, and varied A from 0 to 200 with step size 0.01. For each values of A we calculated U_{Blied}. This plot showed that maximum utility of bad liars occurred at A equals 0 , stating that bad liars should not carry on to lie, instead, they will be better of to report their actual attribute. 119

List of Tables

2.1 Using $X_G \sim N(8,1)$, $X_B \sim N(4,1)$, $L = 10$, $Q = 2$, $p_G = p_B = 0.5$ and $\tau = 6.40$. We obtain the theoretical results of the four probabilities presented in Eq. (2.10) to (2.13), showing the model's predictive power. 22

2.2 The model of Figure 2.1 with $X_G \sim N(8,1)$, $X_B \sim N(4,1)$, $L = 10$, $Q = 2$ and $p_G = p_B = 0.5$. We apply the simulation technique described in Appendix A to obtain the simulation results of the four probabilities presented in Eq. (2.10) to (2.13). The first numerical value shows the average mean values, while the range in the second row entry shows the 95% confidence interval of the average mean values. 23

2.3 In order to satisfy the inequality $\sigma_B^2 - \sigma_G^2 > 0$, we substitute the initial conditions $X_G \sim N(8,1)$, $X_B \sim N(4,2)$, $L = 10$, $Q = 2$ and $p_G = p_B = 0.5$ into Eq. (2.22) to (2.25) to obtain the values of the four probabilities, showing the model's predictive power. 26

2.4 The model of Figure 2.2 with $X_G \sim N(8,1)$, $X_B \sim N(4,2)$, $L = 10$, $Q = 2$ and $p_G = p_B = 0.5$. We apply the simulation technique described in Appendix A to obtain the simulated results of the four probabilities presented in Eq. (2.22) to (2.25). The first numerical value shows the average mean values, while the range in the second row entry shows the 95% confidence interval of the average mean values. 26

2.5 To satisfy the inequality $\sigma_B^2 - \sigma_G^2 < 0$, we substitute $X_G \sim N(8,2)$, $X_B \sim N(4,1)$, $L = 10$, $Q = 2$ and $p_G = p_B = 0.5$ into Eq. (2.34) to (2.37) to obtain the predictive power of the model. 29

2.6 The model of Figure 2.2 with $X_G \sim N(8,2)$, $X_B \sim N(4,1)$, $L = 10$, $Q = 2$ and $p_G = p_B = 0.5$. We apply the simulation technique described in Appendix A to obtain the simulated results of the four probabilities presented in Eq. (2.34) to (2.37). The first numerical value shows the average mean values, while the range in the second row entry shows the standard error interval of the average mean values. 29

2.7 With initial conditions $X_{1G} \sim N(8,1), X_{1B} \sim N(4,1), X_{2G} \sim N(8,1), X_{2B} \sim N(4,1), p_B = p_G = 0.5, L = 10, Q = 2, R = 15.2, E = 9.7$, and varies ρ from -0.9 to 0.9 with increment 0.1. We obtain the four probabilities stated in Eq. (2.52) to (2.55). Note that "ErrorBD" is calculated using Eq. (2.51) and is approximately equals to 0 in different values of ρ, consistent with our theory that we want the total bounded error to be very close to 0. "Sum" represents $P_{BB} + P_{BG} + P_{GB} + P_{GG}$ which equals to 1, satisfies the sum of the four probabilities has to be 1. "$P_{GB} + P_{BG}$" represents the misclassification rate of the model. 37

2.8 Applying the simulation technique in Appendix A and with initial conditions $X_{1G} \sim N(8,1), X_{1B} \sim N(4,1), X_{2G} \sim N(8,1), X_{2B} \sim N(4,1), p_B = p_G = 0.5, L = 10$ and $Q = 2$. We vary ρ from -0.9 to 0.9 with increment 0.1 and apply the set condition (2.39) to classify the group of good payers. We simulate 100 datasets each with 1000 borrowers and rerun the program 100 times to obtain the average mean values of P_{BB}, P_{BG}, P_{GB} and P_{GG}. "Sum" represents $P_{BB} + P_{BG} + P_{GB} + P_{GG}$ which equals to 1, satisfies the theory that the four probabilities have to sum up to 1. "$P_{GB} + P_{BG}$" represents the misclassification rate of the model and is quite low for all values of ρ. 38

2.9 This table shows the standard deviation of the mean values of Table 2.8. Applying the same initial conditions $X_{1G} \sim N(8,1)$, $X_{1B} \sim N(4,1)$, $X_{2G} \sim N(8,1)$, $X_{2B} \sim N(4,1)$, $p_B = p_G = 0.5$, $L = 10$, $Q = 2$ and ρ varies from -0.9 to 0.9 with increment 0.1. We simulate 100 datasets each with 1000 borrowers and rerun the program for 100 times to obtain the standard deviation of the mean values of P_{BB}, P_{BG}, P_{GB} and P_{GG}. 38

2.10 Applying the simulation technique in Appendix A and with initial conditions $X_{1G} \sim N(8,1), X_{1B} \sim N(4,1), X_{2G} \sim N(8,1), X_{2B} \sim N(4,1), p_B = p_G = 0.5, L = 10$ and $Q = 2$. We vary ρ from -0.9 to 0.9 with increment 0.1 and apply the set condition (2.39) to classify the group of good payers. We simulate 100 datasets each with 1000 borrowers and rerun the program 100 times. We sorted the 100 mean values of P_{BB}, P_{BG}, P_{GB} and P_{GG}. This table shows the 5^{th} mean value, which is the lower endpoint of the 95% confidence interval. 39

2.11 Applying the simulation technique in Appendix A and with initial conditions $X_{1G} \sim N(8,1), X_{1B} \sim N(4,1), X_{2G} \sim N(8,1), X_{2B} \sim N(4,1), p_B = p_G = 0.5, L = 10$ and $Q = 2$. We vary ρ from -0.9 to 0.9 with increment 0.1 and apply the set condition (2.39) to classify the group of good payers. We simulate 100 datasets each with 1000 borrowers and rerun the program 100 times. We sorted the 100 mean values of P_{BB}, P_{BG}, P_{GB} and P_{GG}. This table shows the 95^{th} mean value, which is the upper endpoint of the 95% confidence interval. 39

3.1 Parameters used in the case study described in Section 3.4 47

3.2 Both normality tests provide p-value bigger than 0.05. There is not enough evidence to reject the null hypothesis and conclude that X_{2Bnew} is normally distributed. 49

3.3 Using the parameters in Table 3.1 and the simulation method described in Section 3.3. The results are generated using model 1, 2 and 3 are presented above. The first values shows the average value of the 100 datasets, the value after the "±" sign shows the standard error of the average value. 50

4.1 Parameters used in Sections 4.4, 4.5 and 4.6. 61

4.2 Comparison of results generated using different values of A for $L = \$7.2\text{K}$ and each unit of lies costs \$130K to remove. 69

4.3 Comparison of results generated using different values of A for $L = \$7.2\text{K}$ and each unit of lies costs \$260K to remove. 70

4.4 Comparison of the bank's revenue and profit on the two different choices of loan amount: (a) $L = \$7.2\text{K}$ and (b) $L = \$X_i \times 1{,}000$. 72

4.5 A two by two table used to calculate odds ratio to determine whether bad borrowers should lie in order to increase their chance of getting loans. 76

5.1 Parameters used to generate results for each of the six steps described in the introduction 100

5.2 Description of the objective functions that will be presented in the four scenarios mentioned in Section 5.5.2. 104

5.3 Table of results for Scenario 1 where it is not economical to lie. 105

5.4 Table of results for Scenario 2a where it is economical to lie but not to check for lies. 108

5.5 Table of results for Scenario 2b where it is always economical to lie and check for lies. 113

5.6 Table of results for Scenario 2c where it is economical to lie and check for lies periodically. 117

List of Appendices

Appendix A Data Simulation 130
Appendix B Lending Club 133
Appendix C Glossary of Terms 135

Chapter 1

Introduction

1.1 Background

Note that banking terms used in this thesis are defined in Appendix C.

1.1.1 Consumer Credit

Economists define consumer credit as "short and intermediate-term credit extended to individuals through regular business channels, usually to finance the purchase of consumer goods and services or to refinance debts incurred for such purposes". The academic study of consumer credit dates back at least as far as 1929 [2]. In simple terms and as described by Anderson [5], credit allows borrowers to buy now and pay later. A loan is a legal contract in which a borrower receives resources or wealth from the creditor and promises to repay the borrowed amount along with interest in the future. Chawla [11] stated that credit can be used to shift day-to-day expenses in the short term, and can improve people's living by allowing them to invest in real assets or education which increase their purchasing power in the long term. For example, people can use credit to buy their home when they are young, which allows them to distribute their debt into a series of payments which they can make as their earnings and equity rises. Chang and Hanna [3] divided consumer credit into installment credit and revolving credit, while Rea [9] further sub-divided each of them into secured and unsecured credit.

Chien and Devaney [4] defined installment credit as closed-end credit, in which the amount borrowed must be repaid in a specific number of equal payments. Installment credit usually involves larger sums of money and we refer to them as loans. Chien and Devaney [4] also defined revolving credit as open-end credit, for which credit has been granted in advance and with a specific credit limit, consumers can use their credit up to their credit limit any time they prefer and do not have to reapply at the time of use. In addition, consumers can also choose

to continue to use the granted credit while payments of previous debts are still in the process of being paid off. Some revolving products allow borrowers to repay the owed amount and interest at a later time, eg. a year after purchases have been made. The amount owed can also be repaid in full or through a series of equal or unequal payments depending on the agreement between the borrower and the credit grantor.

Credit can further be sub-divided into secured and unsecured loans. The major difference between the two is that a secured loan is protected by one or more assets often termed collateral. Debtors have legal possession of the collateral and in cases where borrowers fail to make payments and default on their loan, debtors can take ownership of the collateral, selling it to recover all or part of the debt. If the sale of the collateral does not raise enough money to pay off the debt, lender can obtain a deficiency judgment against the borrower for the remaining amount. This is the reason for John *et al.* [6] to refer to collateral as a protection of the loan, which reduces risk and gives the lender a specific claim on an asset without reducing her general claim against the borrower. Secured loans usually offer lower interest rates and a longer repayment period but may required a longer processing time. Unsecured installment loans, on the other hand, are loans that do not require the pledge of collateral and the lender relies on the borrower's promise to repay the loan. They involve fewer credit checks, easier and faster access to cash, but requires a higher interest rate and involve smaller amounts of money. Wang [7] examined bank lending policies on unsecured loans and suggested that borrowers with no collateral established their credit-worthiness through repeated interaction with banks. In general, there are four types of consumer credits:

1. Secured installment credit
2. Unsecured installment credit
3. Secured revolving credit
4. Unsecured revolving credit

Mortgage loans are typical examples of secured installment loans, where the amount borrowed will be used to purchase a house and the property will be used to pledge as a collateral. Typically, the value of the house must be verified by a legal appraiser which may lengthen the processing time of the loan. Mortgage loans mostly offer at a relatively lower interest rate, very close to prime rate[1], and the repayment period can be up to 25 years or more. Student loans are in the category of unsecured installment loans; lenders believe that borrowers pursuing their

[1]Prime rate is the best interest rate that banks charge their most credit-worthy customers. In Canada, prime rate is set individually by different financial institutions according to the overnight rate set by the Bank of Canada.

post-secondary education are creditworthy and do not require them to pledge any collateral to secure their loans. Some financial institutions have unsecured personal loan products with interest rate starting at 29.99% (EasyFinancial Canada, July 2016) [2], more than ten times higher than Canada's current prime rate of 2.7% (Bank of Canada, July 2016) [3]. The high interest rate charged on unsecured loans compensates the risk of the lender. In case of a default, the lender does not have liens on any of the borrower's assets and the process of collecting debt on unsecured loans are very complicated and time consuming. Other examples of unsecured installment loans includes unsecured personal loans and payday loans.

Home equity lines of credit are an example of secured revolving loan. Typically, the house of the borrower is pledged as a collateral in which case the loan is called a second lien mortgage. Since the house is a consumer's most valuable asset, homeowners should use home equity lines of credit for major items such as home renovations and medical bills. Consumers have flexible terms of repayment, for example, they can choose to repay interest only and keep the loan opened for as long as they want. A bank credit card is an example of unsecured revolving credit. The credit card application can be done through the internet and the approval process can be as fast as a few minutes. Consumers will be granted with a credit limit which they can use for day-to-day purchases. They do not have to pay off the credit card debt right away, instead, as described by Furletti [10], they can choose to continue to use the credit card as long as they make a minimum monthly payment and stay within their assigned credit limit. However, the interest charged on overdue payments on credit cards are usually very high, ranging from 5% to 25% as of July 2016, depending on the agreement between the issuer and the borrower.

There has been an enormous amount of growth in the consumer credit industry in the past 35 years. Credit will be converted to a debt as soon as the consumer accepted and used the granted credit. We examine the growth of consumer credit in the form of debts. Total household debt is defined as the sum of mortgage debt and consumer debt. Mortgage debt includes loans on all residences and real estate property. Consumer debt includes debt on outstanding credit cards, personal and home equity lines of credit, loans from banks and unpaid bills. Scott [16] revealed the tremendous growth of consumer credit card debt in the United States, where revolving credit card debt has increased at an average annual rate of almost nine percent over the past ten years. On average each US household has eight credit cards in circulation. These

[2] EasyFinancial Canada has their unsecured installment loan starting with an interst rate of 29.99% as of July 2016.

[3] `http://www.bankofcanada.ca/rates/daily-digest/`

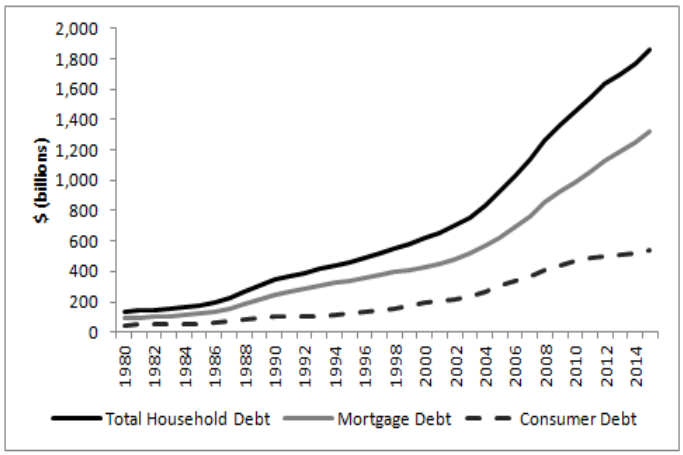


Figure 1.1: This plot shows that Canada's total household debt, mortgage debt and consumer debt have been increasing from 1980 to 2015.

cards were used to purchase nearly $2 trillion of goods and services annually. In Canada, there was a roughly twelve times increase in total household debt between 1980 to 2015. Figure 1.1 shows the plot of Canada total household debt, mortgage debt and consumer debt from 1980 to 2015[4]. Canadian consumer debt increased from $44 billion to $540 billion, while mortgage debt increased from $92 billion to $1,318 billion from 1980 to 2015[5]. To put this into perspective, the Canadian GDP deflator index approximately doubled from 44.24 Index Points in 1980 to 109.92 in 2015. Furthermore, there is a 1.44 times increase in the population in Canada, from 25 million people in 1980 to 36 million in 2015. So, even excluding the effect of inflation and increase in population size, we can still see the dramatic increase in the demand of credit and loans. In 2015, the ratio of household debt to disposable income rose from 85.2% in 1990 to 165.4% in 2015 [6]. This means that as of 2015, Canadian households owed more than $1.65 in debt for every dollar of annual disposable income. Figure 1.2 showed the upward trend of the ratio of Canada household debt to disposable income from 1990 to 2015. The increasing demand for different types of loans urge the development of an automated, low cost, simple and accurate tool, which can effectively and efficiently fulfill the needs of loan request, and can maximize profit from the lending of money. This quantitative tool is called credit scoring.

[4]Sources: Statistics Canada, CANSIM vectors v36408 (total debts), v122698 (consumer debt) and v122736 (mortgage debt), 1980 to 2015.

[5]Numbers provided by Statistics Canada

[6]Sources: Statistics Canada, Table 378-0123, National Balance Sheet Accounts, financial indicators, households and non-profit institutions serving households

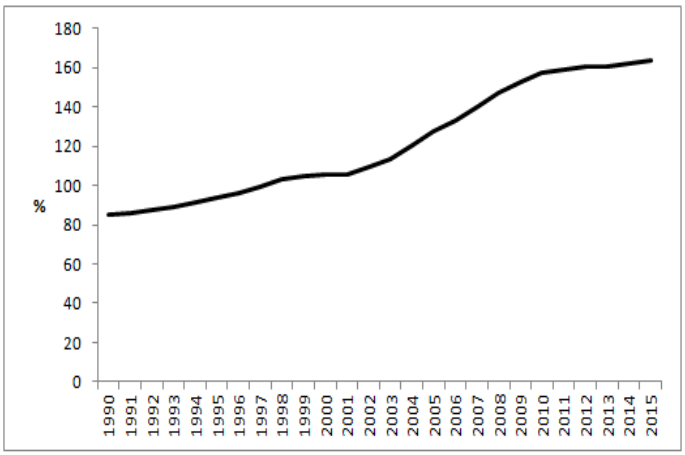


Figure 1.2: This plot shows the ratio of household credit market debt (mortgages plus consumer credit) to disposable income in Canada from 1990 to 2015. The ratio of household debt to disposable income rose from 85.2% in 1990 to 165.4% in 2015. This means that as of 2015, Canada households owed more than $1.65 in debt for every dollar of annual disposable income.

1.1.2 What is Credit Scoring

According to Thomas, Edelman and Crook [21], the idea of credit scoring was introduced more than 60 years ago based on even older statistical ideas. In 1936, Sir Ronald Fisher was the first to identify different groups in a population when one cannot see the characteristic that can clearly defines the groups but only related ones. He segmented two varieties of iris by measuring the physical size of plants; in addition, he also grouped skulls of various origins using their physical measurements. In 1941, David Durand realized that Fisher's work can be further applied to distinguish good and bad loans. This is how credit scoring techniques were first introduced into banking. In 1956, engineer Bill Fair and mathematician Earl Isaac founded the firm Fair, Isaac and Corporation in San Francisco, which was the first credit scoring consultancy company. It is now known as FICO and is still the leading provider of credit analytics and decision management technology in the world. Over the years, FICO and other credit agencies such as Equifax, TransUnion and Experian have convinced the banking community that mathematical formulas and statistical models can do a faster and/or less expensive job of predicting the probability of default of borrowers than most experienced loan officers. The invention of credit cards in the late 1960s magnified the usefulness of credit scoring. The high demand for credit cards, and the relatively small amount borrowed, made it impossible in both economic and manpower terms to handle all the applications without the use of an automated lending system. Financial institutions found that using credit scoring methods to predict credit risk is much better than other judgment based methods and default rates dropped by 50% or more [13]. In the 1980s and 1990s, the successful use of credit scoring in the credit card industry

triggered the application of credit scoring techniques to other financial products, like personal loans, house mortgages and small business loans. Nowadays, the application of credit scoring has become a huge industry which has been widely expanded to many different uses during the last few decades.

According to Thomas *et al.* [21] credit scoring refers to the use of a numerical tool to rank borrowers according to their desirability as loan counterparty. This desirability reflects their probability of making repayments as well as the probability of default (PD), that is the likelihood that a borrower will not be able to meet its debt obligations over a particular time horizon. It also reflects the fraction of loan expected to be recovered in the event of a default, and in opposite terms, the loss given default (LGD) which is the fraction of exposure that cannot be recovered in the event of a default [31]. To build a scoring model, developers require sufficient historical data covering both good and bad economic conditions and reflecting the performance of previously made loans, to determine which borrower characteristics are useful in predicting whether the loan performed well. Data such as the applicant's monthly income, outstanding debt, financial assets, how long the applicant has been in the same job, whether the applicant has defaulted or was ever delinquent on a previous loan, whether the applicant owns or rents a home, and the type of bank account the applicant has are all potential factors that may be related to loan performance, delinquency and default rates of borrowers, and may consider to be used to forecast the probability of default of prospective borrowers. Developers will then come up with a credit scoring model, which is a mathematical formula that can be used to combine the value of the selected characteristics. According to FICO, 50 to 60 variables might be considered when developing a typical scoring model, but 8 to 12 might end up in the final model as yielding the most predictive combination. A credit scoring model will return a credit score for each borrower. For example a common credit score known as the FICO [7] score is a measure of consumer credit risk. This score ranges from 300 to 850; the higher the score the more creditworthy the applicant is deemed to be. As mentioned in Frankel [9], lower credit scores are systematically associated with a higher probability of default on mortgage loans. A good credit scoring model should clearly distinguish between good borrowers who always repay their loans and bad borrowers who default. Therefore, a well-designed model should assign more applicants to high and low scores, while minimizing the number of applicants with scores in the middle range. However, no models are perfect, there will be some good borrowers classified as bad and some bad borrowers classified as good, and a credit scoring model is built to minimize this error. Depending on the amount of risk that the financial institution is willing to take and the amount of return that is expected to obtain, a cutoff score will be set to determine

[7] The acronym FICO is derived from "Fair Issac & Co". This firm is a leading analytics software company which produces statistical models to predict consumer behavior.

whether prospective borrowers will be granted credit.

To apply for a loan, prospective borrowers must provide their attributes by filling out an application form, and also to sign off on a consent form which allows the lender to obtain a copy of their credit history from the credit bureaus. A credit bureau is a company that collects and maintains individual credit information and sells it to lenders, creditors and consumer in the form of a credit report. Chandler and Parker [14] studied the value of credit bureau reports and stated that detailed analysis of the credit report can increase the creditor's ability on predicting risk of loan applicants. With all applicants' information gathered, credit analyst will then input the required attributes into the invented credit scoring model to generate a credit score, and use it to determine the granting decision. Applicants with scores higher than the cutoff score will be granted credit while applicants with sub-cutoff scores will be denied credit. The performance of the credit scoring model must be monitored and validated periodically to make sure that the generated score continue to rank-order borrowers according to their credit quality. Mays [15] stated that change in lending policies, application population and economic conditions are factors that may affect the performance of the scoring model. Different statistical measures such as Kolmogorov Smirnov statistics, probability stability index and characteristic stability analysis have been commonly used to back-test the performance of the credit scoring model and to detect shifts of applicants' characteristics, so as to guarantee that the model continues to return promising granting decisions.

1.1.3 Benefits of Credit Scoring

There are several advantages to both lenders and borrowers of using credit scoring. Credit scoring techniques shorten the processing time for the loan approval processes. The use of high performance computers together with credit scoring systems automated the granting of credits and loans. The use of traditional human-judgment methods to grant such loans can take from a few hours to a few days. After the implementation of credit scoring systems, the granting of small business loans and credit cards fell to as little as a few minutes. In cases where an online application is available, approval of loans or credit can be nearly instantaneous. The introduction of credit scoring reduces the time and cost spent on management. Borrowers can easily have access to the amount of credit they can get from different banks; they can compare and choose the bank which offers the most credit, together with the lowest interest rate and the most suitable repayment terms. The reduction in processing time can minimize the cost and effort to accessing credit, while the manpower saved by both banks and borrowers in processing applications can be effectively allocated to handle more complex issues.

Secondly, credit scoring provides a fair lending process among different borrowers. Under

the United States Equal Credit Opportunity Act [32], variables which reflect gender, race, religion and age cannot be included in credit scoring models, because of their possible link with discrimination against groups of borrowers. Only information that is both non-discriminatory in nature and has been proven to provide predictive power on repayment behavior can be included in the model. Furthermore, since by using credit scoring systems the criteria used to classify good and bad payers have already been set, no human intervention can alter the results of the system. The results of loan requests from different borrowers can be standardized across vast branch networks allowing greater quality control. While human-judgment methods are associated with subjectivity, inconsistency and the common sense of the credit analyst, it is very difficult to apply the same standards to all borrowers. However, a human credit analyst will still take part in the granting of credit if there is evidence that the credit scoring system used is not sufficient to guide the lending decision, and collateral will still be required if the amount of loan is so large that it is questionable whether the customer will be able to repay in full.

Thirdly, credit scoring improves the effectiveness and accuracy of the lending process. Many characteristics are related to the repayment behavior of a borrower but not all of them are important. Credit scoring techniques provide statistical tests to find significant variables which are important in determining the default probability of potential borrowers. Moreover, with the use of computer programs, credit scoring models can be built using larger datasets than a credit analyst can handle and so increase the accuracy of prediction. The invention of high-speed internet service made it possible to request loans from a distance with minimal customer contact; borrowers now have access to many alternative loan products and delivery channels. Also, changes in borrower characteristics can be quickly captured and strategies can be updated accordingly in a fast changing lending environment in which new products are often introduced.

Furthermore, credit scoring helps financial institutions to maximize profit and enabled the development of the sub-prime lending industry. Credit scoring methods have been used to determine the appropriate interest rate that banks should charge different borrowers. High risk borrowers should be charged a higher interest rate compared against low risk borrowers. On the other hand, credit limits can be set differently according to the predicted default probability of potential borrowers. Sub-prime borrowers, who have credit impairment, missing values in their credit histories, (some of the records in their credit histories have no measurements) and difficulties in verifying their income, may be considered to have a higher default risk. As a result, their loan applications may be rejected. The invention of credit scoring allows banks to impose higher interest rate and lower credit limits on sub-prime borrowers. Banks can earn more profit by lending money to sub-prime borrowers despite the occasional loss, meanwhile,

sub-prime borrowers who do not have a strong credit history can still get credit to fulfill their investment or consumption needs.

1.1.4 Other Applications of Credit Scoring

Credit scoring has been applied in areas beyond the lending environment. It techniques have proved to be very useful for targeting customers in direct mailing and other marketing strategies. In the 1950s, Sears [17] used credit scoring techniques to decide whom to send its catalogues. Before sending out a mailing, the name and address of potential customers were prescreened and matched with customers with bad credit histories. Previous bad debtors were filtered out from the mailing group. Moreover, similar techniques as those used to build credit scoring models have been applied to build response models, which can exclude customers who are unlikely to respond and hence reduced the cost by improving the response rate in advertising campaigns. Similar methods can also applied to predict the behaviors of employees; to determine who is likely to be loyal and who is likely to move to a job with another company. A target population may be segmented into several groups and models for predicting which marketing strategy would be better to use on that particular group of customers may be built.

In addition, the preapproval processes on issuing of credit cards and the management of credit card limits can be constructed through the use of credit scoring. According to the Canadian Bankers Association [19], there were 73.9 million credit cards in circulation in Canada in 2012. The high interest rate charged on overdue payments triggered credit card companies to issue preapproved credit cards to customers who have high credit ratings, in order to maximize their profits. Credit card companies regularly review customer's credit limits and amend them according to their credit performance.

Credit scoring has been used by insurance companies to assess risk levels and loss ratios. In the United States, the Circuit Court has found considerable actuarial evidence that credit scores are a good predictor of the risk of loss [18]. A recent actuarial study has concluded that credit scores are one of the most powerful predictors of risk; they are also the most accurate predictor of loss seen in a long time [20]. There is evidence showing that insurance risk and credit scores are strongly correlated with one another. Insurance companies believe that customers with lower credit scores have a higher chance of making insurance claims. According to recent findings, 90% of homeowners and auto insurers use credit scoring to determine insurance premiums and the amount of coverage [21]. The use of credit scoring to predict the risk of loss by insurance companies speeds the process of insurance applications and renewal processes.

1.2 Research Objectives

The main objective of this thesis is to study the effect of fraud towards the lending business. Rezaee [22] defined fraud as a generic term, embracing all multifarious means which human ingenuity can devise, which are resorted to by one individual to get advantage over another by false suggestions, and by suppression of truth. It includes all surprise, trick, cunning, dissembling, and any unfair way by which another is cheated. There are different types of fraud within the banking industry. As stated in Sullivan [23], fraud prevention has become a central concern to banks, customers and public policy makers. Most academic research papers focus on third-party fraud and financial statement fraud. Third-party fraud refers to identity theft in which a person's identity has been stolen to commit crimes. Examples of third-party fraud relate to the use of stolen credit cards and stolen bank accounts. Hoffmann and Birnbrich [30] studied the impact of third-party fraud on the relationship between banks and customers. They showed that customers with better relationships with their banks are less skeptical about anti-fraud measures and thus lower the chance of being an identity fraud victim. Financial statement fraud refers to fraud committed to falsify financial statements, usually committed by management and normally involving overstating income or assets. Rezaee [24] presented fraud prevention and detection strategies in reducing financial statement fraud incidents.

The fundamental theme of this thesis is to focus on the effect of one type of first-party fraud: "bust-out fraud" in the content of consumer lending. Experian [27] explains that bust-out fraud usually involves a long-term planning, fraudsters deliberately borrow and repay money on-time to maintain a good credit history. Additional lines of credit and extended credit limits will be offered to the fraudsters, then eventually, the fraudsters will use all the available credit and abandon the accounts. According to Kitten [25], first-party fraud ranks among the top three fraud events that financial institutions face. In particular, Sumner [28] stated that US card issuers estimate losses from bust-out fraud to be over $1.5 billion annually (Credit Risk International, September 2004). In 2013, FICO analysts [26] estimated that between 10% and 15% of all banks' unsecured bad debt is actually bust-out fraud, resulting in tens of billions in losses every year. It is very difficult to detect this type of fraud and we believe that taking extra precaution on verifying prospective borrowers' characteristics can prevent this type of fraud at the time of loan application. Currently, little effort has been spent on studying whether attributes provided by prospective borrowers should be verified for correctness. Canada's main banking regulator [29], the Office of the Superintendent of Financial Institutions, has recently suggested Canadian banks should verify borrowers' income for mortgage loans. This thesis will look into the effect on consumer lending, in the context of credit scoring, if some borrowers fraudulently falsify their attributes on their loan application form. Note that our methods are

well suited to studying bust-out fraud but are not limited to that.

1.3 Structure of the Thesis

This research comprises both theoretical and numerical investigations in the context of credit scoring on the effect on eliminating fraud. This thesis is composed of six chapters including this Introduction. The main objectives and contents of the subsequent chapters are organized as follows.

Chapter 2 studies the methods used in credit scoring. In particular, we introduce the concept of discriminant analysis, a classification method we employ throughout this thesis. This method has been used in different research literatures to classify borrowers into two groups of good and bad payers. We use simulated data to demonstrate the prediction accuracy of this method. In particular, we show the application of discriminant analysis by assuming the classification model only uses one attribute, which follows a normal distribution to distinguish borrowers. We will further study the prediction power of this method if two attributes both following normal distribution are used in the classification model. In addition, we provide an overview of logistic regression analysis, an alternative method that has been widely applied in credit scoring.

Chapter 3 demonstrates the effect of credit fraud on prediction accuracy. It is reasonable to believe that using larger datasets will improve the predictive power of the resulting credit scoring models. Currently, all credit scoring models are built using as much historical data as possible. We will show that if some borrowers lie in their responses to one or more questions, using higher dimensional data to build credit scoring models may lower the prediction power, compared with using lower dimensional data. This chapter uses simulated data as discussed in Appendix A to show that the improper use of data and models will lead to inaccurate prediction, potentially contributing to big losses.

Chapter 4 further demonstrates that it is profitable for lending institutions to spend extra effort to directly eliminate lies in the dataset. Applicants learn about the characteristics that are used to build credit scoring models, and may have a strong motivation to alter the answers on their application form to improve their chance of loan approval. We develop a credit scoring model which takes into account the potential for borrowers to falsify information. The optimal amount of effort that lenders should spend on identifying liars will be examined to balance risk and return. In addition, we show that fraudsters will eventually educate themselves to adjust their lies, escape from credit checks and obtain loans. The proposed issue will be studied using simulated data as discussed in Appendix A. This chapter can help lending financial institutions to reduce risk and maximize profit, and it also shows that it is feasible for customers to lie

intelligently so as to evade credit checks and get loans.

In Chapter 5, we model the consumer loan approval process as a game between the lender and the borrower. Particularly, we explore the cost for fraudsters to make a clever lie on their attributes and the cost to verify borrowers' characteristics towards different behaviors of lenders and borrowers. We present the equilibrium state of the economy between the lender and the borrower. We show that having successful fraudsters in the loan mix not only affect banks' profitability, but also impact the utility of good borrowers who always repay their loans on time. This research will be studied using discriminant analysis and assume credit scores of borrowers follow a particular class of half triangular distribution. This chapter shows the importance of enriching data before making loan decisions and can help financial institutions to reduce risk and maximize profit.

A summary of the findings of this thesis is presented in Chapter 6. Business insights as well as possible future work motivated by this research is also discussed.

Bibliography

[1] L.J. Mester, What's the point of credit scoring?, *Business Review*, **3**, (1997), pp. 3-16.

[2] L. Calder, *Financing the American dream: a Cultural History of Consumer Credit*, Princeton University Press, Princeton, New Jersey, 2009.

[3] Y.C.R. Chang and S. Hanna, Consumer Credit Search Behavior. *Journal of Consumer Studies & Home Economics*, **16**(3)(1992), 207-227.

[4] Y.W. Chien and S.A. Devaney, The Effects of Credit Attitude and Socioeconomic Factors on Credit Card and Installment Debt, *Journal of Consumer Affairs*, **35**(1), (2001), 162-179, DOI: 10.1111/j.1745-6606.2001.tb00107.x.

[5] R. Anderson, *The Credit Scoring Toolkit: Theory and practice for retail credit risk management and decision automation*, Oxford University Press, Oxford, UK, 2007.

[6] K. John, A. W. Lynch and M. Puri, Credit Ratings, Collateral, and Loan Characteristics: Implications for Yield, *The Journal of Business*, **76**(3), (2003), 371-409.

[7] T. Wang, Paying back to borrow more: reputation and bank credit access in early America, *Explorations in Economic History*, **45**(4), (2008), 477-488.

[8] L.A. Drozd and J.B. Nosal, Competing for customers: A search model of the market for unsecured credit, *Unpublished Manuscript*, University of Wisconsin.

[9] S.A. Rea, Arm-Breaking, Consumer Credit and Personal Bankruptcy, *Economic Inquiry*, **22**(2), (1984), 188-208.

[10] M.J. Furletti, An overview of credit card asset-backed securities, *Available at SSRN 927489* (2002).

[11] R.K. Chawla, The distribution of mortgage debt in Canada. *Perspectives on Labor and Income*, (**23**(2) (2011), 25.

[12] A. Frankel, Prime or Not So Prime? An Exploration of US Housing Finance in the New Century, *BIS Quarterly Review*, (2006), pp. 67-78.

[13] J.H. Myers and E.W. Forgy, The development of numerical credit evaluation systems, *Journal of American Statistical Association*, **58** (1963), 449-470.

[14] G.G. Chandler and L.E. Parker, Predictive value of credit bureau reports, *Journal of Retail Banking*, **11**(4) (1989).

[15] E. Mays, *Handbook of Credit Scoring*, Global Professional Publish, Chicago, USA, 2001.

[16] R.H. Scott, Credit Card Use and Abuse: A Veblenian Analysis, *Journal of Economic Issues*, **41**(2) (2007), 567-574.

[17] M.L. Edward, *An Introduction to Credit Scoring*, Athena Press, London, UK, 1994.

[18] C. Johnson, Abstracts of significant cases bearing on the regulation of insurance, *Journal of Insurance Regulation*, **23** (2005), 81-84.

[19] Canadian Bankers Association, *Credit Cards: Statistics and Facts*, `http://www.cba.ca/en/media-room/50-backgrounders-on-banking-issues/123-credit-cards`

[20] M. Miller, Research confirms value of credit scoring, *National Underwriter*, **107**(42) (2003): 30.

[21] L.C. Thomas, D.B. Edelman and J.N. Crook, *Credit Scoring and Its Applications*, The Society for Industrial and Applied Mathematics, Philadelphia, USA, 2002.

[22] Z. Rezaee, *Financial statement fraud: prevention and detection*, John Wiley & Sons, New York, USA, 2002.

[23] R.J. Sullivan, The changing nature of U.S. card payment fraud: industry and public policy options, *Economic Review*, **95**(2) (2010), pp. 101-33.

[24] Z. Rezaee, Causes, consequences, and deterence of financial statement fraud, *Critical Perspectives on Accounting*, **16**(3) (2005), pp. 277-298.

[25] T. Kitten, First-Party Fraud a Growing Risk, *Experian Research*, `http://www.bankinfosecurity.com/first-party-fraud-growing-risk-a-3836`, (2011-07).

[26] Fair Issac Corporation, *Uncovering Bust-Out Fraud with FICO Identity Resolution Engine*, USA, http://www.fico.com/en/wp-content/secure_upload/Uncovering_Bust_Out_Fraud_with_FICO_Identity_Resolution_Engine_3004WP.pdf, 2013.

[27] Experian plc, *Bust-out fraud: Knowing what to look for can safeguard the bottom line*, USA, http://www.experian.com/assets/decision-analytics/white-papers/bust-out-fraud-white-paper.pdf, 2009.

[28] A. Sumner, *Tackling the Issue of Bust-Out Fraud*, Retail Banker International (2007): 4.

[29] M. Kim and P. Vieira, Canada Banking Regulator Tightens Mortgage-Lending Oversight, *The Wall Street Journal*, http://www.wsj.com/articles/canada-banking-regulator-tightens-mortgage-lending-oversight-1467907512, (2016-07)

[30] A. O. Hoffmann and C. Birnbrich, The impact of fraud prevention on bank-customer relationships: An empirical investigation in retail banking, *International Journal of Bank Marketing*, **30**(5) (2012): 390-407.

[31] T. Schuermann, What do we know about Loss Given Default?, *Credit Risk Models and Management*, **2**(1) (2004).

[32] D. C. Hsia, Credit Scoring and the Equal Credit Opportunity Act, *Hastings LJ*, **30** (1978), pp. 371.

Chapter 2

Methods used in Credit Scoring

In this chapter, we introduce and explain in detail the statistical methods used in this thesis. Different statistical methods have been applied in the building of credit scoring models. According to Lee *et al.* [1], the objective of credit scoring models is to assign credit applicants to either a 'good credit' group that is likely to repay financial obligation or a 'bad credit' group whose application will be denied because of its high probability of defaulting on the financial obligation. Bravo *et al.* [2] further separate the group of 'bad credit' into a group who do not pay because of cash flow problems, and another group that do not pay simply because they are unwilling to pay. Characteristics such as the applicant's monthly income, outstanding debt, financial assets, how long the applicant has been in the same job, whether the applicant has defaulted or was ever delinquent on a previous loan, whether the applicant owns or rents a home, and the types of bank account the applicant has are all potential factors that may be included in the building of credit scoring models.

There are several advantages of applying statistical methods to build credit scoring models. We can use hypothesis testing to identify and incorporate characteristics which reflects repayment behavior, while removing unimportant characteristics from the model. The use of regression analysis shows the relative importance of the characteristics and assigns weights accordingly. The Gini coefficient and receiver operating characteristic (ROC) curve are two popular methods which show the discriminating power of the model. With the use of different statistical techniques, a lean and accurate credit scoring model can be built efficiently.

All the credit scoring models in this thesis will be developed based on the idea of discriminant analysis. Section 2.1 will state the background and examine the predictive power of discriminant analysis. In Section 2.1.1, we assume that applicants may be divided into two groups of good versus bad borrowers using only a single, normally distributed, characteristic. In Section 2.1.2, we extend the assumption to use two characteristics which follow bivariate normal distribution to determine the granting decisions. We will explore in detail the method

used and show the accuracy of discriminant analysis. Section 2.2 gives a brief introduction on the use of logistic regression analysis. Conclusions are drawn in Section 2.3.

2.1 Discriminant Analysis

Discriminant analysis is a well-known classification technique which has been applied to corporate bankruptcies by Altman [4] since 1968. According to Thomas, Edelman and Crook [7] and Thomas [5], the objective of discriminant analysis is to find a model that classifies all borrowers into two subsets, A_G and A_B, representing the groups of good and bad payers respectively. In some literature, for example Steenackers and Goovaerts [6], borrowers are classified into three different groups, the group of good borrowers, bad borrowers and a third indeterminate category. Borrowers in the indeterminate group require further information for their loan approval and will eventually be separated into two groups of good and bad payers. The group of good payers refers to all the borrowers who repay their loans on time, while the group of bad payers has slightly different definitions in different literature. Avery *et al.* [7] define bad payers as those delinquent for more than 60 days, while the Basel Committee on Banking Supervision [8] defines defaulters as those who have any overdue payment of 90 days or more.

In this thesis, we classify all the borrowers into the two categories of good and bad payers, and define the group of bad payers as those who have at least one missed payment. The idea of using discriminant analysis to build credit scoring models is to maximize the expected profit by minimizing the expected credit loss. A model will be built using past payers with known repayment performance, and will be used to determine the granting decisions of new borrowers. Let $Pred$ and Obs be the binary random variable which represents the model's predicted result and the observed result of historical borrower respectively. Let G represent good payers and B represent bad payers. We are interested in finding four different probabilities:

$$P(\{Pred = G\} \cap \{Obs = G\}), \tag{2.1}$$

$$P(\{Pred = G\} \cap \{Obs = B\}), \tag{2.2}$$

$$P(\{Pred = B\} \cap \{Obs = G\}), \tag{2.3}$$

$$P(\{Pred = B\} \cap \{Obs = B\}), \tag{2.4}$$

to reflect the predictive power of the model.

Notice that there is an inherent bias in this approach. Only borrowers who have been granted credit will be considered in the sample used to build models. There is no information about the performance of the applicants whose loan request has been rejected. In other words,

the sample of borrowers used to build a model is representative of those accepted for credit in the past but does not represent all the applicants who applied for credit. According to Hand and Henley [9], this distortion of the distribution of applicants clearly has implications for the accuracy and general applicability of any new credit scoring model that is constructed. Methods used to handle this problem are termed as reject inference. Hand [10] suggests to compare the characteristics of rejected applicants with borrowers whose repayment outcome are known and conclude the outcome of the rejected applicants accordingly. Anderson *et al.* [11] propose three different reject inference methods for which a weight variable will be created and used to assign the good or bad groups to rejected applicants. In real situations, some financial institutions will purchase data from the credit bureaus [1], for which the repayment behavior of rejected applicants with another lending company will be shared and used as a reference to predict their repayment performance as if credit was granted.

A vector of variables $\mathbf{X} = (X_1, X_2, ..., X_p)$, representing the application characteristics of each past payer, will be used to construct the model. In credit-scoring terminology, the actual value of the characteristics for a particular borrower is denoted as $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ and are called the attributes of the characteristics. Attributes are the answer to the application form questions, and characteristics are the questions that were asked, i.e., $\mathbf{X}$ is a random variable and $\mathbf{x}$ is the realization of the random variable. Suppose A is the set of all possible attributes that the application variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ can take, i.e., all the different ways the application form can be answered. The objective is to find a rule that separate the set A into two groups of A_G and A_B, representing two groups of good and bad payers correspondingly.

We assumed that p_G is the proportion of applicants who are good payers, and p_B the proportion of applicants who are bad payers, and these quantities are unknown to the lender. We further assume that $\mathbf{X}$ has a finite number of continuous attributes. Let $f(\mathbf{x}|G)\mathbf{dx}$ be the probability density that a good borrower will report an attribute between $\mathbf{x}$ and $\mathbf{x} + \mathbf{dx}$, and $f(\mathbf{x}|B)\mathbf{dx}$ be the probability that a bad borrower will report an attribute between $\mathbf{x}$ and $\mathbf{dx}$. Notice that applicants' reported attributes may not have the same value as their true attributes, for instance, there may be some noise in the reported data, either due to human error or to applicants purposely altering their attributes to increase their chances of loan approval. In this chapter, we will assume all applicants' reported attributes to be the same as their true attributes. Later in this thesis, we will further distinguish the differences between these two types of attributes and study their effects towards the lending business. Two types of costs correspond to the two types of errors that can be made in this classification method. The first type of error occurs if a good payer was classified as a bad payer. In this case, the potential profit that the lender might have

[1] Credit bureaus are agencies that collects information from various sources and provides consumer credit information on individual consumers for a variety of uses.

been able to earn is lost. The second type of error occurs if a bad payer was classified as a good payer. In this case, losses will be incurred when the borrower defaults on her loan. We assume that the expected misclassification profit and loss incurred, denoted as Q and L respectively, are the same for all borrowers. The purpose of this credit scoring model is to determine the cutoff score used to make granting decisions; borrowers with credit score higher than the cutoff score will be granted credit, otherwise declined.

Again, the intention of the lender is to maximize the expected profit by minimizing the expected credit loss. If a borrower was classified into group A_G, then there is only a loss if the borrower is a bad payer, and the expected loss per borrower is $LP(\mathbf{x}|B)p_B$. On the other hand, if a borrower was put into A_B, there is only a loss if she a good payer, and the corresponding loss per borrower is $QP(\mathbf{x}|G)p_G$. Therefore, we should classify a borrower into A_G if $LP(\mathbf{x}|B)p_B \leq QP(\mathbf{x}|G)p_G$. The solution of this inequality leads to the set:

$$\begin{aligned} A_G &= \{\mathbf{x}|LP(\mathbf{x}|B)p_B \leq QP(\mathbf{x}|G)p_G\} \\ &= \left\{\mathbf{x}\middle|\frac{Lp_B}{Qp_G} \leq \frac{P(\mathbf{x}|G)}{P(\mathbf{x}|B)}\right\} \\ &= \left\{\mathbf{x}\middle|\frac{Lp_B}{Qp_G} \leq \frac{f(\mathbf{x}|G)}{f(\mathbf{x}|B)}\right\}, \end{aligned} \tag{2.5}$$

where the last equality follows if $\mathbf{x}$ is a vector of continuous characteristic variables, and $f(\mathbf{x}|G)$ and $f(\mathbf{x}|B)$ are the probability distributions of $\mathbf{X}_G$ and $\mathbf{X}_B$ respectively. If we can find the distribution of $f(\mathbf{x}|G)$ and $f(\mathbf{x}|B)$, we can then substitute them into the above formula and find the group of good borrowers.

2.1.1 Univariate Normal

Consider the case where there is only one continuous characteristic X, drawn from a normal distribution. The probability density function for good and bad payers are $f(x|G)$ and $f(x|B)$ respectively, where $X_G \sim N(\mu_G, \sigma_G)$ and $X_B \sim N(\mu_B, \sigma_B)$. Thus,

$$f(x|G) = \frac{1}{\sqrt{2\pi}\sigma_G}\exp\left\{\frac{-(x-\mu_G)^2}{2\sigma_G^2}\right\} \text{ and } f(x|B) = \frac{1}{\sqrt{2\pi}\sigma_B}\exp\left\{\frac{-(x-\mu_B)^2}{2\sigma_B^2}\right\}.$$

Applying set condition (2.5), the right hand side of the inequality becomes

$$\frac{f(x|G)}{f(x|B)} = \frac{\frac{1}{\sigma_G}\exp\left\{\frac{-(x-\mu_G)^2}{2\sigma_G^2}\right\}}{\frac{1}{\sigma_B}\exp\left\{\frac{-(x-\mu_B)^2}{2\sigma_B^2}\right\}} = \frac{\sigma_B}{\sigma_G}\exp\left\{\frac{-1}{2}\left[\frac{(x-\mu_G)^2}{\sigma_G^2} - \frac{(x-\mu_B)^2}{\sigma_B^2}\right]\right\} \overset{\text{set}}{\geq} \frac{Lp_B}{Qp_G}.$$

From the last inequality we can deduce that

$$\begin{aligned} & \exp\left\{\frac{-1}{2}\left[\frac{(x-\mu_G)^2}{\sigma_G^2} - \frac{(x-\mu_B)^2}{\sigma_B^2}\right]\right\} \geq && \frac{Lp_B\sigma_G}{Qp_G\sigma_B} \\ \Longrightarrow \quad & \frac{-1}{2}\left[\frac{(x-\mu_G)^2}{\sigma_G^2} - \frac{(x-\mu_B)^2}{\sigma_B^2}\right] \geq && \log\left(\frac{Lp_B\sigma_G}{Qp_G\sigma_B}\right) \\ \Longrightarrow \quad & \frac{(x-\mu_G)^2}{\sigma_G^2} - \frac{(x-\mu_B)^2}{\sigma_B^2} \leq && -2\log\left(\frac{Lp_B\sigma_G}{Qp_G\sigma_B}\right). \end{aligned} \tag{2.6}$$

[Here $\log(x)$ denotes the natural logarithm.]

The model will therefore recommend the bank to grant loans to applicants with attributes in the set A_G, where

$$A_G = \left\{x \middle| \frac{(x-\mu_G)^2}{\sigma_G^2} - \frac{(x-\mu_B)^2}{\sigma_B^2} \leq -2\log\left(\frac{Lp_B\sigma_G}{Qp_G\sigma_B}\right)\right\}. \tag{2.7}$$

Furthermore, different values of σ_G and σ_B will generate different cutoff points. We will consider three different scenarios based on the relationship of σ_G and σ_B. In each scenario, we graphically show the group of good and bad payers, the cutoff points, and the regions where the borrowers will be classified as good payers. We theoretically calculate the four probabilities presented in Eq. (2.1) to (2.4), which represents the predictive power of the model. In addition, we apply the simulation method described in Appendix A and provide a comparison between the theoretical and simulated results.

Case I : $\sigma_{\mathbf{G}}^{\mathbf{2}} = \sigma_{\mathbf{B}}^{\mathbf{2}} = \sigma^{\mathbf{2}}$

If we consider the special case where $\sigma_G^2 = \sigma_B^2 = \sigma^2$, the set condition (2.7) becomes (after some algebra)

$$A_G = \left\{x \middle| x \geq \frac{1}{(\mu_G - \mu_B)}\left[\frac{\mu_G^2 - \mu_B^2}{2} + \sigma^2\log\left(\frac{Lp_B}{Qp_G}\right)\right]\right\}. \tag{2.8}$$

The set condition (2.8) implies the cutoff point is a vertical line with cutoff value

$$\tau = \frac{1}{(\mu_G - \mu_B)}\left[\frac{\mu_G^2 - \mu_B^2}{2} + \sigma^2 \log\left(\frac{Lp_B}{Qp_G}\right)\right], \tag{2.9}$$

as depicted in Figure 2.1. Notice that when setting $\sigma_G = \sigma_B = \sigma$, the x^2 term in Eq. (2.6) canceled out, leading to only one cutoff line. Figure 2.1 shows the probability density function (PDF) of $X_G \sim N(8,1)$ displayed in the solid line, and the PDF of $X_B \sim N(4,1)$ displayed as the dashed line. Substituting the corresponding mean and standard deviation of X_G and X_B, together with $L = 10$, $Q = 2$ and $p_G = p_B = 0.5$ into Eq. (2.9) gives the cutoff value $\tau = 6.40$. The shaded region shows the characteristic values for which customers will be classified as good payers by set condition (2.8). The theoretical calculation of the four probabilities are:

$$P(\{Pred = G\} \cap \{Obs = G\}) = p_G \int_{\tau}^{\infty} f_{x_G}(x)\,dx \quad = p_G P(X_G \geq \tau) = p_G\left[1 - \Phi\left(\frac{\tau - \mu_G}{\sigma_G}\right)\right], \tag{2.10}$$

$$P(\{Pred = G\} \cap \{Obs = B\}) = p_B \int_{\tau}^{\infty} f_{x_B}(x)\,dx \quad = p_B P(X_B \geq \tau) = p_B\left[1 - \Phi\left(\frac{\tau - \mu_B}{\sigma_B}\right)\right], \tag{2.11}$$

$$P(\{Pred = B\} \cap \{Obs = G\}) = p_G \int_{-\infty}^{\tau} f_{x_G}(x)\,dx \quad = p_G P(X_G \leq \tau) = p_G \Phi\left(\frac{\tau - \mu_G}{\sigma_G}\right), \tag{2.12}$$

$$P(\{Pred = B\} \cap \{Obs = B\}) = p_B \int_{-\infty}^{\tau} f_{x_B}(x)\,dx \quad = p_B P(X_B \leq \tau) = p_B \Phi\left(\frac{\tau - \mu_B}{\sigma_B}\right), \tag{2.13}$$

where Φ is the cumulative standard normal distribution function. Setting the initial conditions as $\mu_G = 8, \mu_B = 4, \sigma_G = \sigma_B = 1$, $L = 10$, $Q = 2$, $p_G = p_B = 0.5$ and $\tau = 6.40$, we use Eq. (2.10) to (2.13) to obtain the numerical results of the four probabilities that are stated in Table 2.1.

Furthermore, we applied the simulation method described in Appendix A to simulate 100 datasets each with 500 borrowers and reran the program 100 times. We classified the group of good payers using set condition (2.8) and calculated the average mean values of the four probabilities stated in Eq. (2.1) to (2.4). For each of the probabilities, we computed and sorted the 100 average probabilities. Then, we found the 5^{th} and 95^{th} average probability, which gave the 95% confidence interval of the mean average probability. The results are presented in Table 2.2. The first numerical value showed the average mean values (Avm), and the second row entry showed the 95% confidence interval of each of these mean values. Finally, comparing the values in Table 2.2 and 2.1, the probability calculated theoretically is inside the 95% confidence interval obtained using simulation. We can conclude that both the theoretical and simulation methods give similar results, so are consistent with each other.

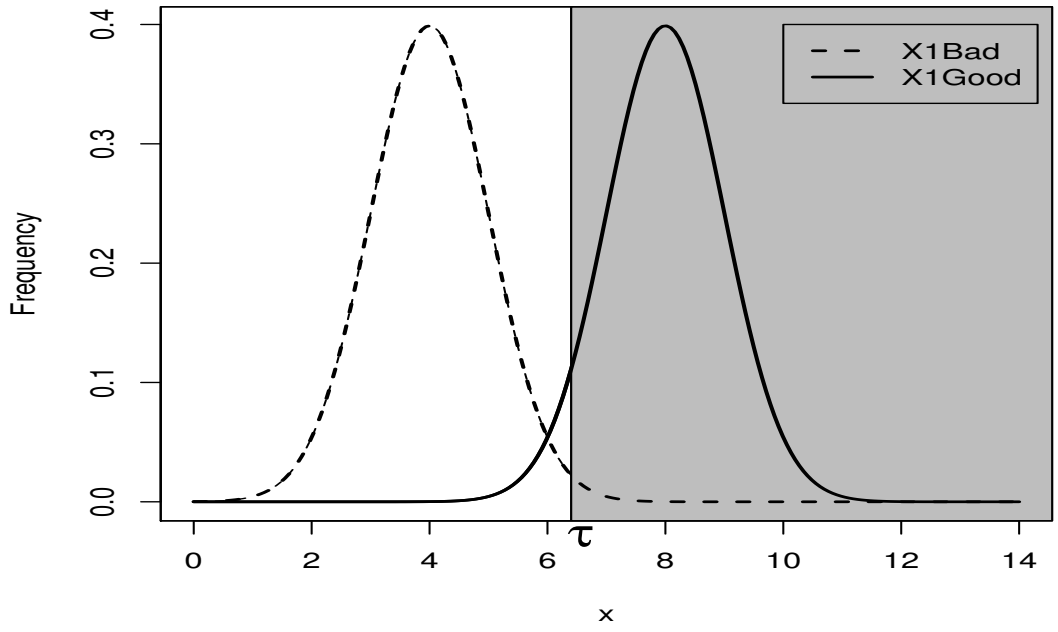


Figure 2.1: Data were simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = P_B = 0.5$. The solid line showing the probability density function (PDF) of X_G and dotted line showing the PDF of X_B. The vertical line shows the cutoff value $\tau = 6.40$ obtained by putting $L = 10$, $Q = 2$, $\mu_G = 8$, $\mu_B = 4$ and $\sigma_G = \sigma_B = \sigma = 1$ into Eq. (2.9). The shaded region shows all the borrowers with characteristics $X_1 > \tau$, and will be classified as good payers by set condition (2.8).

		Observed (*Obs*)	
		Bad	Good
Prediction (*Pred*)	Bad	0.4959	0.0275
	Good	0.0041	0.4725

Table 2.1: Using $X_G \sim N(8,1)$, $X_B \sim N(4,1)$, $L = 10$, $Q = 2$, $p_G = p_B = 0.5$ and $\tau = 6.40$. We obtain the theoretical results of the four probabilities presented in Eq. (2.10) to (2.13), showing the model's predictive power.

Case II : $\sigma_{\mathbf{B}}^{\mathbf{2}} - \sigma_{\mathbf{G}}^{\mathbf{2}} > \mathbf{0}$

Consider the case where $\sigma_B^2 - \sigma_G^2 > 0$. In this case the set condition (2.7) becomes

$$A_G = \{x | b \geq x \geq a\}, \tag{2.14}$$

where

		Observed (*Obs*)	
		Bad	Good
Prediction (*Pred*)	Bad	$\hat{\mu} = 0.4919$ (0.4840, 0.5000)	$\hat{\mu} = 0.0283$ (0.0253, 0.0301)
	Good	$\hat{\mu} = 0.0040$ (0.0031, 0.0047)	$\hat{\mu} = 0.4758$ (0.4667, 0.4837)

Table 2.2: The model of Figure 2.1 with $X_G \sim N(8,1)$, $X_B \sim N(4,1)$, $L = 10$, $Q = 2$ and $p_G = p_B = 0.5$. We apply the simulation technique described in Appendix A to obtain the simulation results of the four probabilities presented in Eq. (2.10) to (2.13). The first numerical value shows the average mean values, while the range in the second row entry shows the 95% confidence interval of the average mean values.

$$b = \sqrt{\frac{1}{\sigma_B^2 - \sigma_G^2}\left[-2\sigma_G^2\sigma_B^2 \log\left(\frac{Lp_B\sigma_G}{Qp_G\sigma_B}\right) - \sigma_B^2\mu_G^2 + \sigma_G^2\mu_B^2\right] + \left(\frac{\mu_G\sigma_B^2 - \mu_B\sigma_G^2}{\sigma_B^2 - \sigma_G^2}\right)} + \frac{\mu_G\sigma_B^2 - \mu_B\sigma_G^2}{\sigma_B^2 - \sigma_G^2} \tag{2.15}$$

and

$$a = -\sqrt{\frac{1}{\sigma_B^2 - \sigma_G^2}\left[-2\sigma_G^2\sigma_B^2 \log\left(\frac{Lp_B\sigma_G}{Qp_G\sigma_B}\right) - \sigma_B^2\mu_G^2 + \sigma_G^2\mu_B^2\right] + \left(\frac{\mu_G\sigma_B^2 - \mu_B\sigma_G^2}{\sigma_B^2 - \sigma_G^2}\right)} + \frac{\mu_G\sigma_B^2 - \mu_B\sigma_G^2}{\sigma_B^2 - \sigma_G^2}. \tag{2.16}$$

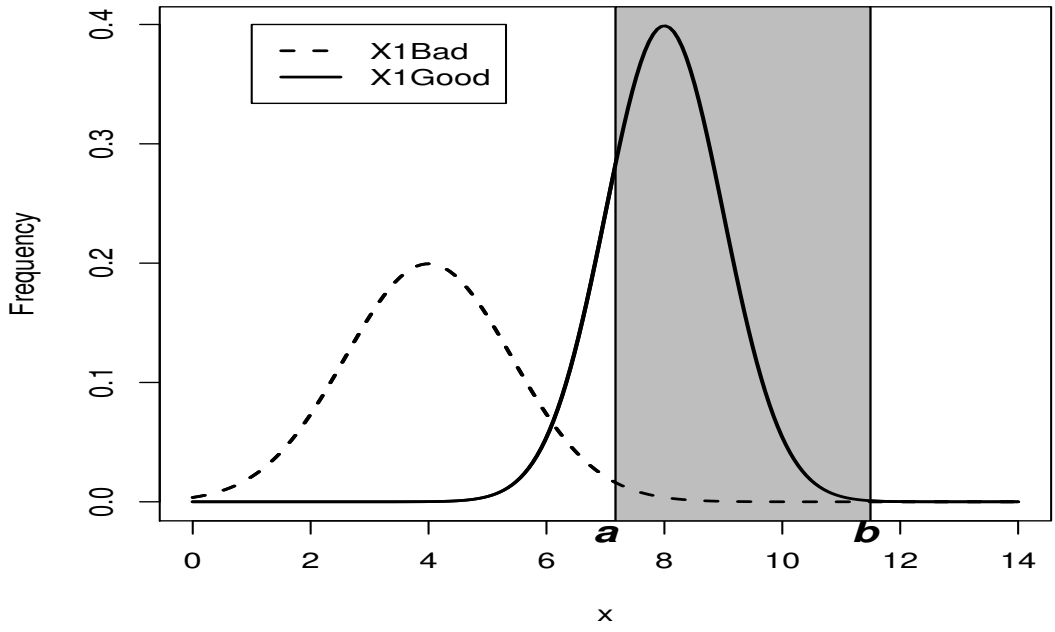


Figure 2.2: Data were simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,2)$ with $P_G = P_B = 0.5$. The solid line showed the probability distribution function (PDF) of X_G, while the dotted line showed the PDF of X_B. The cutoff values were obtained by putting $L = 10$, $Q = 2$, $\mu_G = 8$, $\mu_B = 4$, $\sigma_G = 1$ and $\sigma_B = 2$ into Eq. (2.15) and (2.16). Note that when $\sigma_B^2 > \sigma_G^2$, there were two cutoff points correspond to two vertical cutoff lines with values $a = 7.17$ and $b = 11.49$. The shaded region showed all the borrowers with characteristic values between a and b and was classified as good payers by set condition (2.14).

In order to satisfy the inequality $\sigma_B^2 - \sigma_G^2 > 0$, we changed the initial conditions to $X_G \sim N(8,1)$, $X_B \sim N(4,2)$, $L = 10$, $Q = 2$ and $p_G = p_B = 0.5$. After that, we substitute the initial values into Eq. (2.16) and (2.15) to find the cutoff values a and b to be 7.17 and 11.49 respectively. In the event when $\sigma_G > \sigma_B$, the model will classify borrowers with characteristic values between a and b as good payers. Conversely, the model will classify borrowers with characteristic values outside the range of a and b as bad payers. This suggest two cutoff lines as displayed in Figure 2.2. In addition, we know that

$$\mu_G + 4\sigma_G = 8 + 4 \times 1 = 12 \qquad \text{and} \qquad \mu_B + 4\sigma_B = 4 + 4 \times 2 = 12. \tag{2.17}$$

However, for all values of $K > 4$

$$\mu_B + K\sigma_B > \mu_G + K\sigma_G. \tag{2.18}$$

In particular,

$$\mu_G + 5\sigma_G = 8 + 5 \times 1 = 13 \qquad \text{and} \qquad \mu_B + 4.5\sigma_B = 4 + 4.5 \times 2 = 13. \tag{2.19}$$

From the common properties of normal distribution, more observations that fall within 5σ of the mean than within 4.5σ of the mean. In other words, the normal tail for bad payers contains more observations than the tail representing good payers when $x > 12$. This contributes to having two regions of bad borrowers as shown in Figure 2.2. This is a very interesting fact that applicants with attribute value that is "too good to be true" will be classified as bad borrowers. However, this is more of a theoretical curiosity than a practical outcome. Indeed, the proportion of good payers with attribute value greater than 11.49 is

$$p_G P(X_G \geq 11.49) = p_G(1 - \Phi\left(\frac{11.49 - 8}{1}\right)) = 0.00012 \text{ (Rare)}. \tag{2.20}$$

and the proportion of bad payers with attribute value greater than 11.49 is

$$p_B P(X_B \geq 11.49) = p_B(1 - \Phi\left(\frac{11.49 - 4}{1}\right)) = 1.72 \times 10^{-14} \approx 0 \text{ (Rare)}. \tag{2.21}$$

We can conclude that there is a very low probability that applicants will have an attribute value greater than 11.49. Again, we are interested in finding the four different probabilities to determine the accuracy of the model. Under the given condition,

$$P(\{Pred = G\} \cap \{Obs = G\}) = p_G \int_a^b f_{x_G}(x)\,dx = p_G\left[\Phi\left(\frac{b - \mu_G}{\sigma_G}\right) - \Phi\left(\frac{a - \mu_G}{\sigma_G}\right)\right], \tag{2.22}$$

$$P(\{Pred = G\} \cap \{Obs = B\}) = p_B \int_a^b f_{x_B}(x)\,dx = p_B\left[\Phi\left(\frac{b - \mu_B}{\sigma_B}\right) - \Phi\left(\frac{a - \mu_B}{\sigma_B}\right)\right], \tag{2.23}$$

$$\begin{aligned} P(\{Pred = B\} \cap \{Obs = G\}) &= p_G\left[\int_b^\infty f_{x_G}(x)\,dx + \int_{-\infty}^a f_{x_G}(x)\,dx\right] \\ &= p_G\left[1 - \Phi\left(\frac{b - \mu_G}{\sigma_G}\right) + \Phi\left(\frac{a - \mu_G}{\sigma_G}\right)\right], \end{aligned} \tag{2.24}$$

$$\begin{aligned} P(\{Pred = B\} \cap \{Obs = B\}) &= p_B\left[\int_{-\infty}^a f_{x_B}(x)\,dx + \int_b^{-\infty} f_{x_B}(x)\,dx\right] \\ &= p_B\left[1 - \Phi\left(\frac{b - \mu_B}{\sigma_B}\right) + \Phi\left(\frac{a - \mu_B}{\sigma_B}\right)\right], \end{aligned} \tag{2.25}$$

where Φ is the cumulative standard normal distribution.

We substitute the initial values into Eq. (2.22) to (2.25) to compute the four probabilities,

and present the findings in Table 2.3. Note that the $P(\{Pred = B\} \cap \{Obs = G\}) \approx 0.1$, although quite high, is consistent with Figure 2.2, which is the area under the solid line in the non-shaded region. This means that many good payers are misclassified as bad under the given condition.

		Observed (*Obs*)	
		Bad	Good
Prediction (*Pred*)	Bad	0.4719	0.1022
	Good	0.0281	0.3978

Table 2.3: In order to satisfy the inequality $\sigma_B^2 - \sigma_G^2 > 0$, we substitute the initial conditions $X_G \sim N(8,1)$, $X_B \sim N(4,2)$, $L = 10$, $Q = 2$ and $p_G = p_B = 0.5$ into Eq. (2.22) to (2.25) to obtain the values of the four probabilities, showing the model's predictive power.

Moreover, we applied the simulation method described in Appendix A to simulate 100 datasets each with 500 borrowers and reran the program 100 times. We used set condition (2.14) to classify the group of good payers, obtained the simulated results of average mean values of the four probabilities, calculated the 95% confidence interval and presented the results in Table 2.4. By comparing the values in Table 2.3 and Table 2.4, we can conclude that the simulated results are consistent with the theoretical results.

		Observed (*Obs*)	
		Bad	Good
Prediction (*Pred*)	Bad	$\hat{\mu} = 0.4669$ (0.4591, 0.4754)	$\hat{\mu} = 0.1041$ (0.0995, 0.1094)
	Good	$\hat{\mu} = 0.0290$ (0.0267, 0.0315)	$\hat{\mu} = 0.4000$ (0.3880, 04095)

Table 2.4: The model of Figure 2.2 with $X_G \sim N(8,1)$, $X_B \sim N(4,2)$, $L = 10$, $Q = 2$ and $p_G = p_B = 0.5$. We apply the simulation technique described in Appendix A to obtain the simulated results of the four probabilities presented in Eq. (2.22) to (2.25). The first numerical value shows the average mean values, while the range in the second row entry shows the 95% confidence interval of the average mean values.

Case III : $\sigma_{\mathbf{B}}^{\mathbf{2}} - \sigma_{\mathbf{G}}^{\mathbf{2}} < \mathbf{0}$

Consider the case where $\sigma_B^2 - \sigma_G^2 < 0$. Here the set condition (2.7) becomes:

$$A_G = \{x | x \geq b \text{ or } x \leq a\}, \tag{2.26}$$

where

$$b = \sqrt{\frac{1}{\sigma_B^2 - \sigma_G^2}\left[-2\sigma_G^2\sigma_B^2 \log\left(\frac{Lp_B\sigma_G}{Qp_G\sigma_B}\right) - \sigma_B^2\mu_G^2 + \sigma_G^2\mu_B^2\right] + \left(\frac{\mu_G\sigma_B^2 - \mu_B\sigma_G^2}{\sigma_B^2 - \sigma_G^2}\right)} + \frac{\mu_G\sigma_B^2 - \mu_B\sigma_G^2}{\sigma_B^2 - \sigma_G^2} \tag{2.27}$$

and

$$a = -\sqrt{\frac{1}{\sigma_B^2 - \sigma_G^2}\left[-2\sigma_G^2\sigma_B^2 \log\left(\frac{Lp_B\sigma_G}{Qp_G\sigma_B}\right) - \sigma_B^2\mu_G^2 + \sigma_G^2\mu_B^2\right] + \left(\frac{\mu_G\sigma_B^2 - \mu_B\sigma_G^2}{\sigma_B^2 - \sigma_G^2}\right)} + \frac{\mu_G\sigma_B^2 - \mu_B\sigma_G^2}{\sigma_B^2 - \sigma_G^2}. \tag{2.28}$$

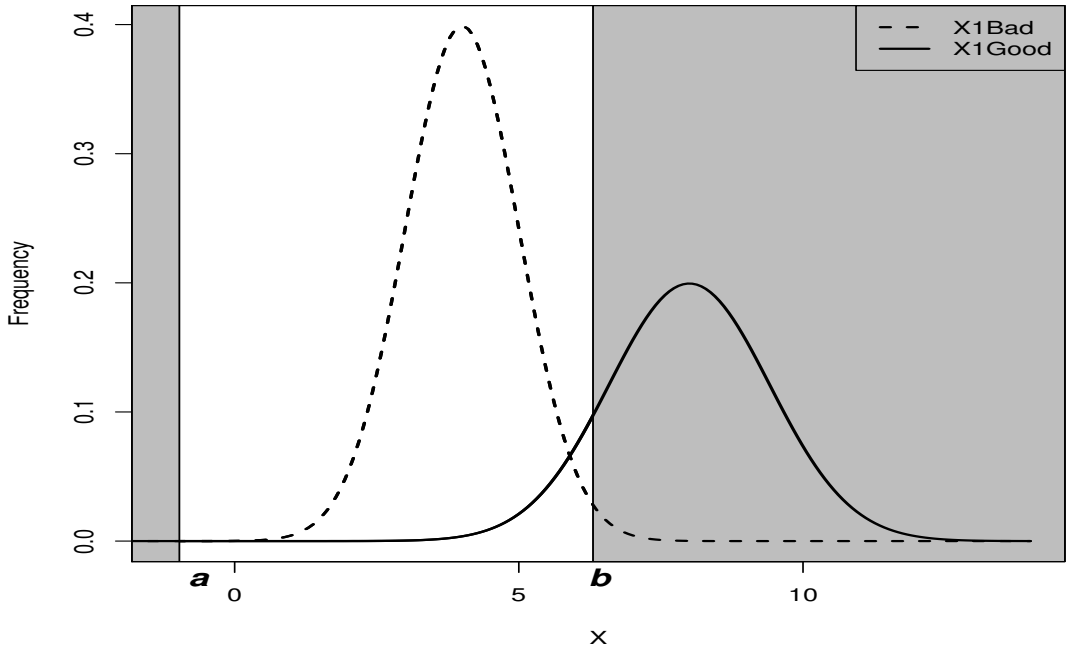


Figure 2.3: Data were simulated from $X_G \sim N(8,2)$ and $X_B \sim N(4,1)$ with $P_G = P_B = 0.5$. The solid line displayed the probability distribution function (PDF) of X_G, while the dotted line displayed the PDF of X_B. The cutoff values were obtained by putting $L = 10$, $Q = 2$, $\mu_G = 8$, $\mu_B = 4$, $\sigma_G = 2$ and $\sigma_B = 1$ into Eq. (2.27) and (2.28). Note that when $\sigma_G^2 > \sigma_B^2$, there were two cutoff points with values $a = -0.97$ and $b = 6.31$. The shaded region showed the borrowers with characteristics that will be classified as good payers by set condition (2.26).

In order to satisfy the inequality $\sigma_B^2 - \sigma_G^2 < 0$, we changed the initial conditions to $X_G \sim N(8,2)$, $X_B \sim N(4,1)$, $L = 10$, $Q = 2$ and $p_G = p_B = 0.5$. Using Eq. (2.27) and (2.28), we calculated the cutoff values a and b to be -0.97 and 6.31 respectively. In the event when $\sigma_G < \sigma_B$, the model classified borrowers with characteristic values between a and b as bad payers, while borrowers having characteristic values outside the region as good. This suggest

two cutoff lines as shown in Figure 2.3. In addition, we know that

$$\mu_G - 4\sigma_G = 8 - 4\times 2 = 0 \qquad \text{and} \qquad \mu_B - 4\sigma_B = 4 - 4\times 1 = 0. \tag{2.29}$$

However, for all values of $K < 4$

$$\mu_B + K\sigma_B > \mu_G + K\sigma_G. \tag{2.30}$$

In particular,

$$\mu_G - 5\sigma_G = 8 - 5\times 2 = -2 \qquad \text{and} \qquad \mu_B - 6\sigma_B = 4 - 6\times 1 = -2. \tag{2.31}$$

From the common properties of normal distribution, more observations fall within 6σ of the mean than within 5σ of the mean. In other words, the normal tail represented by bad payers contains fewer observations than the tail represented by good payers when $x < -0.97$. This implies that there are two regions of good borrowers as depicted in Figure 2.3. Note that the proportion of good and bad payers with attribute value less than -0.97 is

$$p_B P(X_B \le -0.97) = 0.5\Phi\left(\frac{-0.97-4}{1}\right) = 1.67\times 10^{-5} \text{ (Rare)} \tag{2.32}$$

and

$$p_G P(X_G \le -0.97) = 0.5\Phi\left(\frac{-0.97-8}{1}\right) = 1.48\times 10^{-19} \approx 0 \text{ (Rare)} \tag{2.33}$$

respectively. Again, we are interested in finding the four different probabilities to determine the accuracy of the model. Under the given condition,

$$\begin{aligned} P(\{Pred = G\}\cap\{Obs = G\}) &= p_G\left[\int_{-\infty}^{a} f_{x_G}(x)\,dx + \int_{b}^{\infty} f_{x_G}(x)\,dx\right] \\ &= p_G\left[\Phi\left(\frac{a-\mu_G}{\sigma_G}\right) + 1 - \Phi\left(\frac{b-\mu_G}{\sigma_G}\right)\right], \end{aligned} \tag{2.34}$$

$$\begin{aligned} P(\{Pred = G\}\cap\{Obs = B\}) &= p_B\left[\int_{-\infty}^{a} f_{x_B}(x)\,dx + \int_{b}^{\infty} f_{x_B}(x)\,dx\right] \\ &= p_B\left[\Phi\left(\frac{a-\mu_B}{\sigma_B}\right) + 1 - \Phi\left(\frac{b-\mu_B}{\sigma_B}\right)\right], \end{aligned} \tag{2.35}$$

$$P(\{Pred = B\}\cap\{Obs = G\}) = p_G\int_{a}^{b} f_{x_G}(x)\,dx = p_G\left[\Phi\left(\frac{b-\mu_G}{\sigma_G}\right) - \Phi\left(\frac{a-\mu_G}{\sigma_G}\right)\right], \tag{2.36}$$

$$P(\{Pred = B\}\cap\{Obs = B\}) = p_B\int_{a}^{b} f_{x_B}(x)\,dx = p_B\left[\Phi\left(\frac{b-\mu_B}{\sigma_B}\right) - \Phi\left(\frac{a-\mu_B}{\sigma_B}\right)\right], \tag{2.37}$$

where Φ is the cumulative standard normal distribution function.

We graphically showed the cutoff lines, the probability distribution function of X_G and X_B, and the region where the model will classify borrowers as good payers in Figure 2.3. On the other hand, we show the model's predictive power by inserting the initial values into Eq. (2.34) to (2.37), and stated the results of the four probabilities in Table 2.5. Notice that $P(\{Pred = B\} \cap \{Obs = G\}) \approx 0.1$, although quite high, is consistent with Figure 2.2, which is the area under the solid line in the non-shaded region.

		Observed (*Obs*)	
		Bad	Good
Prediction (*Pred*)	Bad	0.4947	0.0993
	Good	0.0053	0.4007

Table 2.5: To satisfy the inequality $\sigma_B^2 - \sigma_G^2 < 0$, we substitute $X_G \sim N(8,2)$, $X_B \sim N(4,1)$, $L = 10$, $Q = 2$ and $p_G = p_B = 0.5$ into Eq. (2.34) to (2.37) to obtain the predictive power of the model.

Again, we applied the simulation technique introduced in Appendix A to simulate 100 datasets each with 500 borrowers. We reran the program for 100 times and used set condition (2.26) to classify the group of good payers. We presented the simulated results and calculated the 95% confidence intervals in Table 2.6. By comparing the values in Table 2.5 and Table 2.6, we can conclude that the simulated results are consistent with the theoretical results.

		Observed (*Obs*)	
		Bad	Good
Prediction (*Pred*)	Bad	$\hat{\mu} = 0.4906$ (0.4829, 0.4984)	$\hat{\mu} = 0.1015$ (0.0968, 0.1070)
	Good	$\hat{\mu} = 0.0053$ (0.0042, 0.0062)	$\hat{\mu} = 0.4026$ (0.3917, 0.4121)

Table 2.6: The model of Figure 2.2 with $X_G \sim N(8,2)$, $X_B \sim N(4,1)$, $L = 10$, $Q = 2$ and $p_G = p_B = 0.5$. We apply the simulation technique described in Appendix A to obtain the simulated results of the four probabilities presented in Eq. (2.34) to (2.37). The first numerical value shows the average mean values, while the range in the second row entry shows the standard error interval of the average mean values.

2.1.2 Bivariate Normal

Now, consider another case where there are two continuous normally distributed characteristics X_1 and X_2, where $X_{1G} \sim N(\mu_{1G}, \sigma_{1G})$, $X_{1B} \sim N(\mu_{1B}, \sigma_{1B})$, $X_{2G} \sim N(\mu_{2G}, \sigma_{2G})$, $X_{2B} \sim$

$N(\mu_{2B}, \sigma_{2B})$ and $corr(X_1, X_2) = \rho$. Therefore, the variance covariance matrix of the good and bad payers are:

$$\Sigma_G = \begin{pmatrix} \sigma_{1G}^2 & \rho\sigma_{1G}\sigma_{2G} \\ \rho\sigma_{1G}\sigma_{2G} & \sigma_{2G}^2 \end{pmatrix} \quad \text{and} \quad \Sigma_B = \begin{pmatrix} \sigma_{1B}^2 & \rho\sigma_{1B}\sigma_{2B} \\ \rho\sigma_{1G}\sigma_{2G} & \sigma_{2B}^2 \end{pmatrix}.$$

The corresponding density function for good payers and bad payers are

$$f(\mathbf{x}|G) = \frac{1}{2\pi|\Sigma_G|^{\frac{1}{2}}} \exp\left\{\frac{-1}{2}(\mathbf{x}-\mu_\mathbf{G})^T\Sigma_G^{-1}(\mathbf{x}-\mu_\mathbf{G})\right\}$$

and

$$f(\mathbf{x}|B) = \frac{1}{2\pi|\Sigma_B|^{\frac{1}{2}}} \exp\left\{\frac{-1}{2}(\mathbf{x}-\mu_\mathbf{B})^T\Sigma_B^{-1}(\mathbf{x}-\mu_\mathbf{B})\right\}.$$

Applying set condition (2.5), the right hand side of the inequality becomes

$$\begin{aligned} \frac{f(x|G)}{f(x|B)} &= \frac{\frac{1}{2\pi\sigma_{1G}\sigma_{2G}} \exp\left\{\frac{-1}{2}(\mathbf{x}-\mu_\mathbf{G})^T\Sigma_G^{-1}(\mathbf{x}-\mu_\mathbf{G})\right\}}{\frac{1}{2\pi\sigma_{1B}\sigma_{2B}} \exp\left\{\frac{-1}{2}(\mathbf{x}-\mu_\mathbf{B})^T\Sigma_B^{-1}(\mathbf{x}-\mu_\mathbf{B})\right\}} \\ &= \frac{\sigma_{1B}\sigma_{2B}}{\sigma_{1G}\sigma_{2G}} \exp\left\{\frac{-1}{2}(\mathbf{x}-\mu_\mathbf{G})^T\Sigma_G^{-1}(\mathbf{x}-\mu_\mathbf{G}) - (\mathbf{x}-\mu_\mathbf{B})^T\Sigma_B^{-1}(\mathbf{x}-\mu_\mathbf{B})\right\} \\ &\overset{\text{set}}{\geq} \frac{Lp_B}{Qp_G}. \end{aligned}$$

From the above, we conclude the model showing the set of good payers is

$$A_G = \left\{\mathbf{x} \mid \mathbf{x}^T(\Sigma_G^{-1} - \Sigma_B^{-1})\mathbf{x} - 2\mathbf{x}(\Sigma_G^{-1}\mu_\mathbf{G} - \Sigma_B^{-1}\mu_\mathbf{B}) \leq \mu_\mathbf{B}^\mathbf{T}\Sigma_B^{-1}\mu_\mathbf{B} - \mu_\mathbf{G}^\mathbf{T}\Sigma_G^{-1}\mu_\mathbf{G} - 2\log\left(\frac{\sigma_{1G}\sigma_{2G}Lp_B}{\sigma_{1B}\sigma_{2B}Qp_G}\right)\right\}. \tag{2.38}$$

From the above set inequality, we know that different choices of variance covariance matrix will generate different types of cutoff lines. When the model contains only one characteristic X, drawn from a normal distribution, all the cutoff lines are straight lines. However, when the model contains two characteristics X_1 and X_2, from a bivariate normal, the cutoff lines can be a straight or a curved line depending on the values of the variance covariance matrix of the good and bad payers. We further illustrate the bivariate normal case by assuming good and bad borrowers share the common variance covariance matrix Σ. Then Eq. (2.38) becomes

$$A_G = \left\{\mathbf{x}|\mathbf{x}\,\Sigma^{-1}(\mu_G - \mu_B)^T \geq \frac{\mu_G\Sigma^{-1}\mu_G^T - \mu_G\Sigma^{-1}\mu_B^T}{2} + \log\left(\frac{Lp_B}{Qp_G}\right)\right\}. \tag{2.39}$$

We further assume that

$$\Sigma^{-1} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix},$$

where $\sigma_{12} = \rho\sigma_1\sigma_2$ and we know that $\mathbf{x} = (x_1, x_2)$. Then Eq. (2.39) becomes

$$\alpha x_1 + \beta x_2 \geq \gamma,$$

where

$$\alpha = \frac{\sigma_2^2(\mu_{1G} - \mu_{1B}) - \sigma_{12}(\mu_{2G} - \mu_{2B})}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2},$$

$$\beta = \frac{\sigma_1^2(\mu_{2G} - \mu_{2B}) - \sigma_{12}(\mu_{1G} - \mu_{1B})}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2},$$

and

$$\gamma = \frac{f(\mu_{1G}, \mu_{2G}, \sigma_1, \sigma_2, \sigma_{12}) - f(\mu_{1B}, \mu_{2B}, \sigma_1, \sigma_2, \sigma_{12})}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} + \log\left(\frac{Lp_B}{Qp_G}\right),$$

where

$$f(\mu_{1G}, \mu_{2G}, \sigma_1, \sigma_2, \sigma_{12}) = (\sigma_2^2\mu_{1G} - \sigma_{12}\mu_{2G})\mu_{1G} + (\sigma_1^2\mu_{2G} - \sigma_{12}\mu_{1G})\mu_{2G},$$

and

$$f(\mu_{1B}, \mu_{2B}, \sigma_1, \sigma_2, \sigma_{12}) = (\sigma_2^2\mu_{1B} - \sigma_{12}\mu_{2B})\mu_{1B} + (\sigma_1^2\mu_{2B} - \sigma_{12}\mu_{1B})\mu_{2B}. \tag{2.40}$$

In order to have a clear understanding of the way the cutoff line classifies the set of good borrowers, we conduct a simulation study. We generate random bivariate normal numbers, X_1 and X_2, using the function mvrnorm() in R [3] package **MASS** [12]. We set $\mu_{1G} = \mu_{2G} = 8$, $\mu_{1B} = \mu_{2B} = 4$, $\sigma_{1G} = \sigma_{2G} = \sigma_{1B} = \sigma_{2B} = 1$ and assume the correlation between X_1 and X_2 to be zero. Four different combinations of X_1 and X_2 compose of (X_{1B}, X_{2B}), (X_{1B}, X_{2G}), (X_{1G}, X_{2B}) and (X_{1G}, X_{2G}) are generated. Figure 2.4 shows the contour plots of the probability density functions of each pair of X_1 and X_2. The solid contour on the top right corner represents borrowers with both characteristics taking the "good" values X_{1G} and X_{2G}, the dotted contours on the top left corner represents borrowers with X_{1B} and X_{2G}, the solid contours on the lower left corner represents borrowers with both characteristics taking the "bad" values X_{1B} and X_{2B}, and the dotted contours on the right bottom corner represents borrowers having X_{1G} and X_{2B}. Using set condition (2.40), we know that the cutoff line for the bivariate case is $\alpha x_1 + \beta x_2 = \gamma$, where α, β and γ are as stated in the condition. Substituting the initial values stated above into α, β and γ, together with $L = 10$ and $Q = 2$, the cutoff line becomes

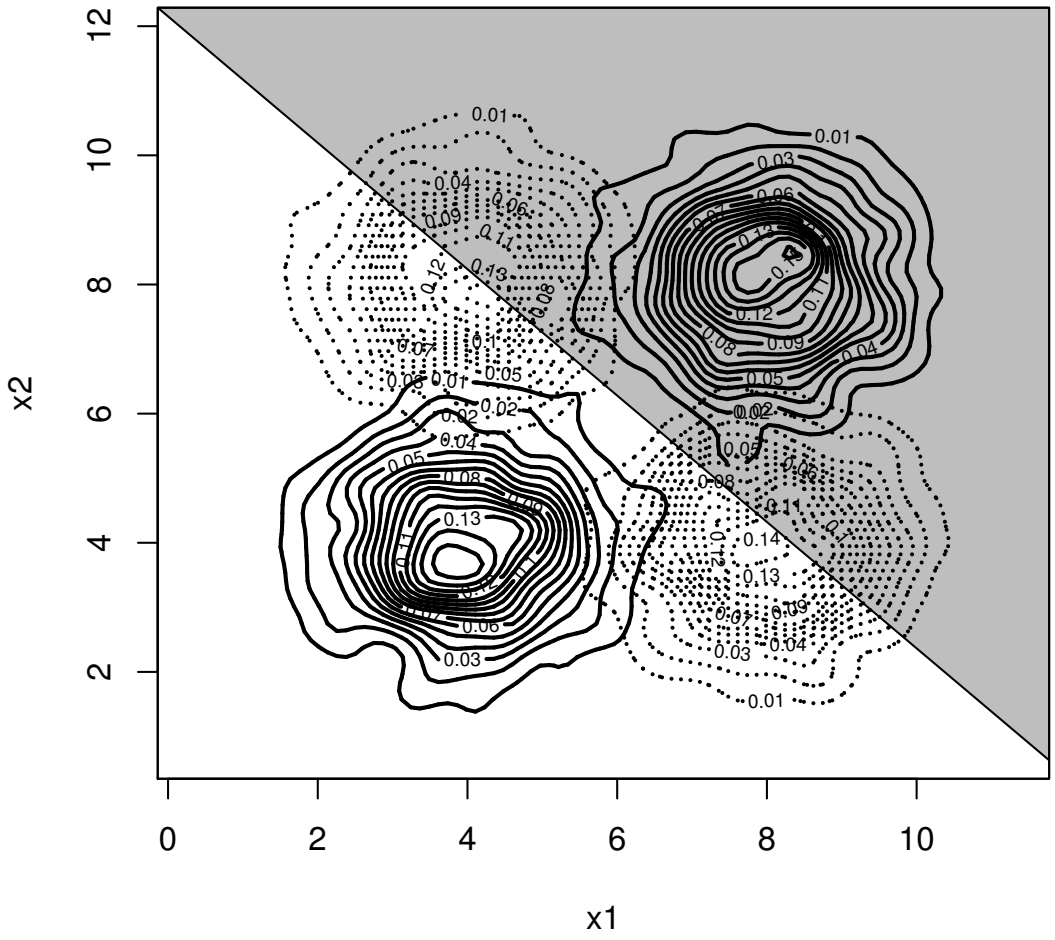


Figure 2.4: Data was simulated using $X_{1G} \sim N(8,1), X_{1B} \sim N(4,1), X_{2G} \sim N(8,1), X_{2B} \sim N(4,1), p_B = p_G = 0.5, L = 10,\ Q = 2$ and $\rho = 0$. The solid contour on the top right corner of the plot represents borrowers with characteristics X_{1G} and X_{2G}, while the solid contour on the lower left corner represents borrowers with characteristics X_{1B} and X_{2B}. The dotted contour on the top left corner represents borrowers with characteristics X_{1B} and X_{2G}, and the dotted contour on the lower right corner represents borrowers with characteristics X_{1G} and X_{2B}. The diagonal line is the cutoff line plotted using Eq. 2.41 and the shaded region shows borrowers having characteristics that will be classified as good payers by the set condition 2.39.

$$x_2 = 12 + \frac{1}{4}\log(\frac{L}{Q}) - x_1, \tag{2.41}$$

which is linear and is shown as the diagonal line in Figure 2.4. The shaded region showed the borrowers with characteristics X_1 and X_2 who will be classified as good payers. In ad-

dition, Figure 2.4 shows the discriminating power of set condition (2.40). The cutoff line in Eq. (2.41) gives accurate classification to borrowers having characteristics (X_{1B}, X_{2B}) and (X_{1G}, X_{2G}), while borrowers having characteristics (X_{1B}, X_{2G}) and (X_{1G}, X_{2B}) will be classified as good payers depending on their actual X_{1G} or X_{2G} values. For example, X_1 may represent the income of the borrower, while X_2 represents the age of the credit history of the borrower. Higher income and longer credit history favors the borrowing of money. The pair of characteristics (X_{1B}, X_{2G}) represents borrowers with low income but long credit history, while the pair of characteristics (X_{1G}, X_{2B}) represents borrowers with high income but short credit history. Whether the borrower will be classified as a good payer depends on how high their income is if they have characteristics (X_{1G}, X_{2B}), and how long they have their credit history if they have characteristics (X_{1B}, X_{2G}).

Again, we are interested in evaluating the four different probabilities in Eq. (2.1) to (2.4) to determine the accuracy of the model. Under the given condition, the four probabilities become:

$$P(\{Pred = G\} \cap \{Obs = G\}) = p_G \int_{-\infty}^{\infty} \int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty} f_{X_{1G},X_{2G}}(x_1\ ,\ x_2)\, dx_1\, dx_2, \tag{2.42}$$

$$P(\{Pred = G\} \cap \{Obs = B\}) = p_B \int_{-\infty}^{\infty} \int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty} f_{X_{1B},X_{2B}}(x_1\ ,\ x_2)\, dx_1\, dx_2, \tag{2.43}$$

$$P(\{Pred = B\} \cap \{Obs = G\}) = p_G \int_{-\infty}^{\infty} \int_{-\infty}^{\frac{\gamma-\beta x_2}{\alpha}} f_{X_{1G},X_{2G}}(x_1\ ,\ x_2)\, dx_1\, dx_2, \tag{2.44}$$

$$P(\{Pred = B\} \cap \{Obs = B\}) = p_B \int_{-\infty}^{\infty} \int_{-\infty}^{\frac{\gamma-\beta x_2}{\alpha}} f_{X_{1B},X_{2B}}(x_1\ ,\ x_2)\, dx_1\, dx_2, \tag{2.45}$$

where $f_{X_1,X_2}(x_1\ ,\ x_2)$ is the probability distribution function of bivariate normal with correlation ρ. In order to evaluate Eq. (2.42) to (2.45), notice that

$$\begin{aligned} &\int_{-\infty}^{\infty} \int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty} f(x_1\ ,\ x_2)\, dx_1\, dx_2 \qquad (2.46)\\ &= \int_{-\infty}^{-E} \int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty} f(x_1\ ,\ x_2)\, dx_1\, dx_2 + \int_{-E}^{R} \int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty} f(x_1\ ,\ x_2)\, dx_1\, dx_2 + \int_{R}^{\infty} \int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty} f(x_1\ ,\ x_2)\, dx_1\, dx_2, \end{aligned}$$

where $R > 0$ and $E > 0$. We want to choose the smallest R and E consistent with

$$\left| \int_{-\infty}^{-E} \int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty} f(x_1\ ,\ x_2)\, dx_1\, dx_2 \right| << \left| \int_{-E}^{R} \int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty} f(x_1\ ,\ x_2)\, dx_1\, dx_2 \right|,$$

and also

$$\left| \int_{R}^{\infty} \int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty} f(x_1\ ,\ x_2)\, dx_1\, dx_2 \right| << \left| \int_{-E}^{R} \int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty} f(x_1\ ,\ x_2)\, dx_1\, dx_2 \right|,$$

then we can state that

$$\underbrace{\int_{-\infty}^{\infty}\int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty} f(x_1\ ,\ x_2)\,dx_1\,dx_2}_{\text{DI}} \approx \underbrace{\int_{-E}^{R}\int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty} f(x_1\ ,\ x_2)\,dx_1\,dx_2}_{\text{AD}}. \tag{2.47}$$

Note that

$$\underbrace{\int_{R}^{\infty}\int_{-\infty}^{\infty} f(x_1\ ,\ x_2)\,dx_1\,dx_2}_{\text{(BB1)}} > \int_{R}^{\infty}\int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty} f(x_1\ ,\ x_2)\,dx_1\,dx_2$$

and

$$\underbrace{\int_{-\infty}^{-E}\int_{-\infty}^{\infty} f(x_1\ ,\ x_2)\,dx_1\,dx_2}_{\text{(BB3)}} > \int_{-\infty}^{-E}\int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty} f(x_1\ ,\ x_2)\,dx_1\,dx_2.$$

We want to find R and E such that (BB1), (BB3) and also the bounded error in evaluating (AD) are small. Then we can apply the corresponding R and E using Eq. (2.47) to approximate the double integral (DI).

To approximate the value of $P(\{Pred = G\} \cap \{Obs = G\})$, we have to first evaluate (BB1) and (BB3). Under the condition that both the predicted and observed values are good, (BB1) becomes

$$C\int_{R}^{\infty}\int_{-\infty}^{\infty}\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_{1G})^2}{\sigma_{1G}^2}+\frac{(x_2-\mu_{2G})^2}{\sigma_{2G}^2}-\frac{2\rho(x_1-\mu_{1G})(x_2-\mu_{2G})}{\sigma_{1G}\sigma_{2G}}\right]\right\}dx_1\,dx_2,$$
$$\textit{where}\ \ C=\frac{1}{2\pi\sigma_{1G}\sigma_{2G}\sqrt{1-\rho^2}}. \tag{2.48}$$

By completing the square in the inner integral and canceling the common terms, Eq. (2.48) equals

$$\frac{1}{\sqrt{2\pi}\sigma_{2G}}\int_{R}^{\infty}\exp\left\{\frac{-(x_2-\mu_{2G})^2}{2\sigma_{2G}^2}\right\}dx_2=\left[1-\Phi\left(\frac{R-\mu_{2G}}{\sigma_{2G}}\right)\right].$$

Similarly, (BB3) becomes

$$\Phi\left(\frac{-E-\mu_{2G}}{\sigma_{2G}}\right). \tag{2.49}$$

Our purpose is to find R and E such that the total bounded error of the four probabilities of interest are all small. Denote the error bound of $P(\{Pred = G\} \cap \{Obs = G\}), P(\{Pred = G\} \cap \{Obs = B\}), P(\{Pred = B\} \cap \{Obs = G\})$ and $P(\{Pred = B\} \cap \{Obs = B\})$ as EBgg, EBgb,

EBbg and EBbb respectively. Using the same technique as above we can conclude that

$$EBgg = p_G\left\{\left[1-\Phi\left(\frac{R-\mu_{2G}}{\sigma_{2G}}\right)\right]+\Phi\left(\frac{-E-\mu_{2G}}{\sigma_{2G}}\right)+\text{Bounded Error in (AD)}\right\}, \tag{2.50}$$
$$EBgb = p_B\left\{\left[1-\Phi\left(\frac{R-\mu_{2B}}{\sigma_{2B}}\right)\right]+\Phi\left(\frac{-E-\mu_{2B}}{\sigma_{2B}}\right)+\text{Bounded Error in (AD)}\right\},$$
$$EBbg = p_G\left\{\left[1-\Phi\left(\frac{R-\mu_{2G}}{\sigma_{2G}}\right)\right]+\Phi\left(\frac{-E-\mu_{2G}}{\sigma_{2G}}\right)+\text{Bounded Error in (AD)}\right\},$$
$$EBbb = p_B\left\{\left[1-\Phi\left(\frac{R-\mu_{2B}}{\sigma_{2B}}\right)\right]+\Phi\left(\frac{-E-\mu_{2B}}{\sigma_{2B}}\right)+\text{Bounded Error in (AD)}\right\}.$$

Therefore,

$$\text{Total Bounded Error}(R,E) = \text{EBgg}+\text{EBgb}+\text{EBbg}+\text{EBbb}. \tag{2.51}$$

In order to obtain the bounded error in (AD) we use the R package **cubature** [13], which evaluates multidimensional integration using adaptive methods over hypercubes. For details on using adaptive algorithms to solve multiple integrals, one can refer to Berntsen *et al.* [14]. There is only one function in the package **cubature**, called adaptIntegrate(), which returns the value of the integral, the estimated relative error and the number of times the function was evaluated. We used the function and equated the bounded error in (AD) as the estimated relative error returned by the function. We put in different values of R and E to calculate the corresponding total bounded error, then we chose the R and E which gave the smallest total bounded error. In addition, we used the same R function and restricted the maximum number of evaluations to 100 to approximate (DI) using (AD). If both the predicted and observed values are good, then by completing the squares and canceling the common terms,

$$\overbrace{\int_{-E}^{R}\int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty} f(x_1\,,\,x_2)\,dx_1\,dx_2}^{\text{AD}}$$
$$= \frac{p_G}{2\pi\sigma_{1G}\sigma_{2G}\sqrt{1-\rho^2}}\int_{-E}^{R}\int_{\frac{\gamma-\beta x_2}{\alpha}}^{\infty}\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_{1G})^2}{\sigma_{1G}^2}+\frac{(x_2-\mu_{2G})^2}{\sigma_{2G}^2}-\frac{2\rho(x_1-\mu_{1G})(x_2-\mu_{2G})}{\sigma_{1G}\sigma_{2G}}\right]\right\}dx_1\,dx_2$$
$$= \frac{p_G}{\sqrt{2\pi}\sigma_{2G}}\int_{-E}^{R}\exp\left\{\frac{-(x_2-\mu_{2G})^2}{2\sigma_{2G}^2}\right\}\left[1-\Phi\left(\frac{\frac{\frac{\gamma-\beta x_2}{\alpha}-\mu_{1G}}{\sigma_{1G}}-\frac{\rho(x_2-\mu_{2G})}{\sigma_{2G}}}{\sqrt{1-\rho^2}}\right)\right]dx_2.$$

Therefore, we can conclude that

$$P(\{Pred=G\}\cap\{Obs=G\}) \approx \frac{p_G}{\sqrt{2\pi}\sigma_{2G}}\int_{-E}^{R}\exp\left\{\frac{-(x_2-\mu_{2G})^2}{2\sigma_{2G}^2}\right\}\left[1-\Phi\left(\frac{\frac{\frac{\gamma-\beta x_2}{\alpha}-\mu_{1G}}{\sigma_{1G}}-\frac{\rho(x_2-\mu_{2G})}{\sigma_{2G}}}{\sqrt{1-\rho^2}}\right)\right]dx_2, \tag{2.52}$$

$$P(\{Pred = G\} \cap \{Obs = B\}) \approx \frac{p_B}{\sqrt{2\pi}\sigma_{2B}} \int_{-E}^{R} \exp\left\{\frac{-(x_2-\mu_{2B})^2}{2\sigma_{2B}^2}\right\} \left[1 - \Phi\left(\frac{\frac{\gamma-\beta x_2}{\alpha} - \mu_{1B} - \frac{\rho(x_2-\mu_{2B})}{\sigma_{2B}}}{\sqrt{1-\rho^2}}\right)\right] dx_2, \tag{2.53}$$

$$P(\{Pred = B\} \cap \{Obs = G\}) \approx \frac{p_G}{\sqrt{2\pi}\sigma_{2G}} \int_{-E}^{R} \exp\left\{\frac{-(x_2-\mu_{2G})^2}{2\sigma_{2G}^2}\right\} \Phi\left(\frac{\frac{\gamma-\beta x_2}{\alpha} - \mu_{1G} - \frac{\rho(x_2-\mu_{2G})}{\sigma_{2G}}}{\sqrt{1-\rho^2}}\right) dx_2, \tag{2.54}$$

$$P(\{Pred = B\} \cap \{Obs = B\}) \approx \frac{p_B}{\sqrt{2\pi}\sigma_{2B}} \int_{-E}^{R} \exp\left\{\frac{-(x_2-\mu_{2B})^2}{2\sigma_{2B}^2}\right\} \Phi\left(\frac{\frac{\gamma-\beta x_2}{\alpha} - \mu_{1B} - \frac{\rho(x_2-\mu_{2B})}{\sigma_{2B}}}{\sqrt{1-\rho^2}}\right) dx_2. \tag{2.55}$$

With initial conditions $X_{1G} \sim N(8,1), X_{1B} \sim N(4,1), X_{2G} \sim N(8,1), X_{2B} \sim N(4,1), p_B = p_G = 0.5, L = 10$ and $Q = 2$, we again used the function adaptIntegrate() and found that the smallest R and E values which give the smallest bounded error to be 15.2 and 9.7 respectively. By setting $R = 15.2$ and $E = 9.7$ in Eq. (2.52) to (2.55) and applying adaptIntegrate() with different ρ values which varies from -0.9 to 0.9 with increment 0.1, we obtained the theoretical results stated in Table 2.7. The row "sum" in Table 2.7 was evaluated using $P_{BB} + P_{BG} + P_{GB} + P_{GG}$ and was equal to 1, satisfy the condition that the four probabilities had to sum to 1. "ErrorBD" was calculated using Eq. (2.51) and is approximately 0 in all the entries, consistent with our theory that the total bounded error has to be close to 0. Furthermore, we used the same initial conditions, simulated both X_1 and X_2 using the method described in Appendix A and applying the set condition in Eq. (2.39) to find the set of good payers. We simulated 100 datasets each with 1000 borrowers and reran the program for 100 times to obtain the average mean values of the four probabilities. The simulated results of the average mean values of the four probabilities are provided in Table 2.8. Table 2.9 shows the standard deviation of the mean values of P_{BB}, P_{BG}, P_{GB} and P_{GG}. In order to obtain the confidence interval estimate of the mean values of the four probabilities for different values of ρ, we sorted the 100 mean probabilities. Then, we found the 5^{th} and 95^{th} mean probability, which gave the 95% confidence interval of the mean probabilities. Table 2.10 shows the lower standard error bars of the mean values of P_{BB}, P_{BG}, P_{GB} and P_{GG}, while Table 2.11 shows the upper standard error bars of the mean values. For each ρ, the theorectical values of the four probabilities as stated in Table 2.7 are within the standard error bars, which provide evidence that simulation and theoretical approach are mutually consistent.

ρ	X_1 Only	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
P_{BB}	0.4959	0.4964	0.4968	0.4972	0.4976	0.4980	0.4984	0.4987	0.4990	0.4993
P_{BG}	0.0275	0.0242	0.0211	0.0180	0.0152	0.0125	0.0100	0.0078	0.0058	0.0041
P_{GB}	0.0041	0.0036	0.0032	0.0028	0.0024	0.0020	0.0016	0.0013	0.0010	0.0007
P_{GG}	0.4725	0.4758	0.4789	0.4820	0.4848	0.4875	0.4900	0.4922	0.4942	0.4959
Sum	1	1	1	1	1	1	1	1	1	1
Error BD		0.0279	0.0192	0.0144	0.0136	0.0135	0.0085	0.0060	0.0049	0.0044
$P_{BG}+P_{GB}$	0.0316	0.0279	0.0243	0.0208	0.0175	0.0144	0.0116	0.0090	0.0067	0.0048
ρ	0	-0.1	-0.2	-0.3	-0.4	-0.5	-0.6	-0.7	-0.8	-0.9
P_{BB}	0.4973	0.4997	0.4998	0.4999	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
P_{BG}	0.0027	0.0017	0.0009	0.0004	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
P_{GB}	0.0005	0.0003	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
P_{GG}	0.4995	0.4983	0.4991	0.4996	0.4998	0.5000	0.5000	0.5000	0.5000	0.5000
Sum	1	1	1	1	1	1	1	1	1	1
Error BD	0.0042	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041
$P_{BG}+P_{GB}$	0.0032	0.0020	0.0011	0.0005	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000

Table 2.7: With initial conditions $X_{1G} \sim N(8,1), X_{1B} \sim N(4,1), X_{2G} \sim N(8,1), X_{2B} \sim N(4,1), p_B = p_G = 0.5, L = 10, Q = 2, R = 15.2, E = 9.7$, and varies ρ from -0.9 to 0.9 with increment 0.1. We obtain the four probabilities stated in Eq. (2.52) to (2.55). Note that "ErrorBD" is calculated using Eq. (2.51) and is approximately equals to 0 in different values of ρ, consistent with our theory that we want the total bounded error to be very close to 0. "Sum" represents $P_{BB}+P_{BG}+P_{GB}+P_{GG}$ which equals to 1, satisfies the sum of the four probabilities has to be 1. "$P_{GB}+P_{BG}$" represents the misclassification rate of the model.

2.2 Logistic Regression

Logistic regression is another most frequently used statistical modeling method in the context of credit scoring. One of the earliest use of logistic regression on credit scoring dates back to 1980, where Wiginton [15] showed that logistic regression yields higher accuracy in predicting consumer credit behavior than discriminant analysis, however, neither method was sufficiently good to be cost effective for the dataset that he employed. Bensic *et al.* [16] compared the classification power of logistic regression, neural networks and CART decision tree on small business loans using a relatively small dataset, they revealed that logistic regression showed higher discriminatory power than decision tree, but not as good as neural networks. Efron [18] compared the efficiency of logistic regression with discriminant analysis and concluded that logistic regression is shown to be between one half or two thirds as effective as discriminant analysis. In addition, Press and Wilson [19] also compared the use of logistic regression with discriminant analysis, they stated that when the data are normally distributed with identical covariance matrices, the prediction accuracy of discriminant analysis is higher than that of

ρ	X_1 Only	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
$\mu_{P_{BB}}$	0.4960	0.4965	0.4969	0.4973	0.4977	0.4982	0.4985	0.4989	0.4992	0.4994
$\mu_{P_{BG}}$	0.0275	0.0241	0.0210	0.0180	0.0152	0.0125	0.0100	0.0078	0.0058	0.0041
$\mu_{P_{GB}}$	0.0041	0.0036	0.0032	0.0028	0.0024	0.0019	0.0016	0.0012	0.0009	0.0007
$\mu_{P_{GG}}$	0.4724	0.4758	0.4789	0.4819	0.4847	0.4874	0.4899	0.4921	0.4941	0.4958
Sum	1	1	1	1	1	1	1	1	1	1
$\mu_{P_{BG}+P_{GB}}$	0.0315	0.0278	0.0242	0.0208	0.0175	0.0144	0.0115	0.0090	0.0067	0.0048
ρ	0	-0.1	-0.2	-0.3	-0.4	-0.5	-0.6	-0.7	-0.8	-0.9
$\mu_{P_{BB}}$	0.4996	0.4998	0.4999	0.5000	0.5001	0.5001	0.5001	0.5001	0.5001	0.5001
$\mu_{P_{BG}}$	0.0028	0.0016	0.0009	0.0004	0.0001	0	0	0	0	0
$\mu_{P_{GB}}$	0.0005	0.0003	0.0002	0.0001	0	0	0	0	0	0
$\mu_{P_{GG}}$	0.4971	0.4983	0.4990	0.4995	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999
Sum	1	1	1	1	1	1	1	1	1	1
$\mu_{P_{BG}+P_{GB}}$	0.0032	0.0019	0.0011	0.0005	0.0002	0	0	0	0	0

Table 2.8: Applying the simulation technique in Appendix A and with initial conditions $X_{1G} \sim N(8,1), X_{1B} \sim N(4,1), X_{2G} \sim N(8,1), X_{2B} \sim N(4,1), p_B = p_G = 0.5, L = 10$ and $Q = 2$. We vary ρ from -0.9 to 0.9 with increment 0.1 and apply the set condition (2.39) to classify the group of good payers. We simulate 100 datasets each with 1000 borrowers and rerun the program 100 times to obtain the average mean values of P_{BB}, P_{BG}, P_{GB} and P_{GG}. "Sum" represents $P_{BB} + P_{BG} + P_{GB} + P_{GG}$ which equals to 1, satisfies the theory that the four probabilities have to sum up to 1. "$P_{GB} + P_{BG}$" represents the misclassification rate of the model and is quite low for all values of ρ.

ρ	X_1 Only	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
$\sigma_{\mu_{P_{BB}}}$	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016
$\sigma_{\mu_{P_{BG}}}$	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0003	0.0003	0.0002	0.0002
$\sigma_{\mu_{P_{GB}}}$	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001
$\sigma_{\mu_{P_{GG}}}$	0.0016	0.0016	0.0015	0.0016	0.0015	0.0016	0.0015	0.0016	0.0016	0.0016
ρ	0	-0.1	-0.2	-0.3	-0.4	-0.5	-0.6	-0.7	-0.8	-0.9
$\sigma_{\mu_{P_{BB}}}$	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016
$\sigma_{\mu_{P_{BG}}}$	0.0002	0.0001	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\sigma_{\mu_{P_{GB}}}$	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\sigma_{\mu_{P_{GG}}}$	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016

Table 2.9: This table shows the standard deviation of the mean values of Table 2.8. Applying the same initial conditions $X_{1G} \sim N(8,1)$, $X_{1B} \sim N(4,1)$, $X_{2G} \sim N(8,1)$, $X_{2B} \sim N(4,1)$, $p_B = p_G = 0.5$, $L = 10$, $Q = 2$ and ρ varies from -0.9 to 0.9 with increment 0.1. We simulate 100 datasets each with 1000 borrowers and rerun the program for 100 times to obtain the standard deviation of the mean values of P_{BB}, P_{BG}, P_{GB} and P_{GG}.

ρ	X_1 only	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
$5^{th}\mu_{P_{BB}}$	0.4933	0.4938	0.4943	0.4946	0.4953	0.4955	0.4958	0.4963	0.4964	0.4968
$5^{th}\mu_{P_{BG}}$	0.0265	0.0233	0.0201	0.0174	0.0146	0.0119	0.0094	0.0072	0.0054	0.0037
$5^{th}\mu_{P_{GB}}$	0.0038	0.0033	0.0030	0.0025	0.0021	0.0017	0.0013	0.0011	0.0008	0.0006
$5^{th}\mu_{P_{GG}}$	0.4695	0.4731	0.4762	0.4792	0.4821	0.4849	0.4873	0.4894	0.4913	0.4930
ρ	0	-0.1	-0.2	-0.3	-0.4	-0.5	-0.6	-0.7	-0.8	-0.9
$5^{th}\mu_{P_{BB}}$	0.4970	0.4971	0.4973	0.4974	0.4974	0.4974	0.4974	0.4974	0.4974	0.4974
$5^{th}\mu_{P_{BG}}$	0.0025	0.0014	0.0007	0.0003	0.0001	0	0	0	0	0
$5^{th}\mu_{P_{GB}}$	0.0004	0.0002	0.0001	0	0	0	0	0	0	0
$5^{th}\mu_{P_{GG}}$	0.4945	0.4955	0.4964	0.4968	0.4971	0.4972	0.4973	0.4973	0.4973	0.4973

Table 2.10: Applying the simulation technique in Appendix A and with initial conditions $X_{1G} \sim N(8,1), X_{1B} \sim N(4,1), X_{2G} \sim N(8,1), X_{2B} \sim N(4,1), p_B = p_G = 0.5, L = 10$ and $Q = 2$. We vary ρ from -0.9 to 0.9 with increment 0.1 and apply the set condition (2.39) to classify the group of good payers. We simulate 100 datasets each with 1000 borrowers and rerun the program 100 times. We sorted the 100 mean values of P_{BB}, P_{BG}, P_{GB} and P_{GG}. This table shows the 5^{th} mean value, which is the lower endpoint of the 95% confidence interval.

ρ	X_1 only	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
$95^{th}\mu_{P_{BB}}$	0.4986	0.4992	0.4996	0.4998	0.5004	0.5008	0.5011	0.5015	0.5019	0.5020
$95^{th}\mu_{P_{BG}}$	0.0283	0.0249	0.0217	0.0186	0.0159	0.0131	0.0105	0.0082	0.0062	0.0045
$95^{th}\mu_{P_{GB}}$	0.0044	0.0040	0.0035	0.0030	0.0026	0.0022	0.0018	0.0014	0.0011	0.0008
$95^{th}\mu_{P_{GG}}$	0.4749	0.4784	0.4814	0.4844	0.4871	0.4900	0.4924	0.4946	0.4966	0.4982
ρ	0	-0.1	-0.2	-0.3	-0.4	-0.5	-0.6	-0.7	-0.8	-0.9
$95^{th}\mu_{P_{BB}}$	0.5021	0.5023	0.5025	0.5026	0.5026	0.5027	0.5027	0.5027	0.5027	0.5027
$95^{th}\mu_{P_{BG}}$	0.0031	0.0019	0.0010	0.0005	0.0002	0.0001	0	0	0	0
$95^{th}\mu_{P_{GB}}$	0.0006	0.0004	0.0002	0.0001	0.0001	0	0	0	0	0
$95^{th}\mu_{P_{GG}}$	0.4997	0.5007	0.5014	0.5020	0.5021	0.5023	0.5023	0.5023	0.5023	0.5023

Table 2.11: Applying the simulation technique in Appendix A and with initial conditions $X_{1G} \sim N(8,1), X_{1B} \sim N(4,1), X_{2G} \sim N(8,1), X_{2B} \sim N(4,1), p_B = p_G = 0.5, L = 10$ and $Q = 2$. We vary ρ from -0.9 to 0.9 with increment 0.1 and apply the set condition (2.39) to classify the group of good payers. We simulate 100 datasets each with 1000 borrowers and rerun the program 100 times. We sorted the 100 mean values of P_{BB}, P_{BG}, P_{GB} and P_{GG}. This table shows the 95^{th} mean value, which is the upper endpoint of the 95% confidence interval.

logistic regression. Different researchers have their own preferences on the modeling method used to study their research. We choose discriminant analysis as our modeling method in this thesis, but we would like to provide an overview of logistic regression, an alternative method that can be use in most classification problems. We will discuss about our preference on discriminant analysis over logistic regression at the end of this section.

Logistic regression is a special type of generalized linear model which allows the use of a

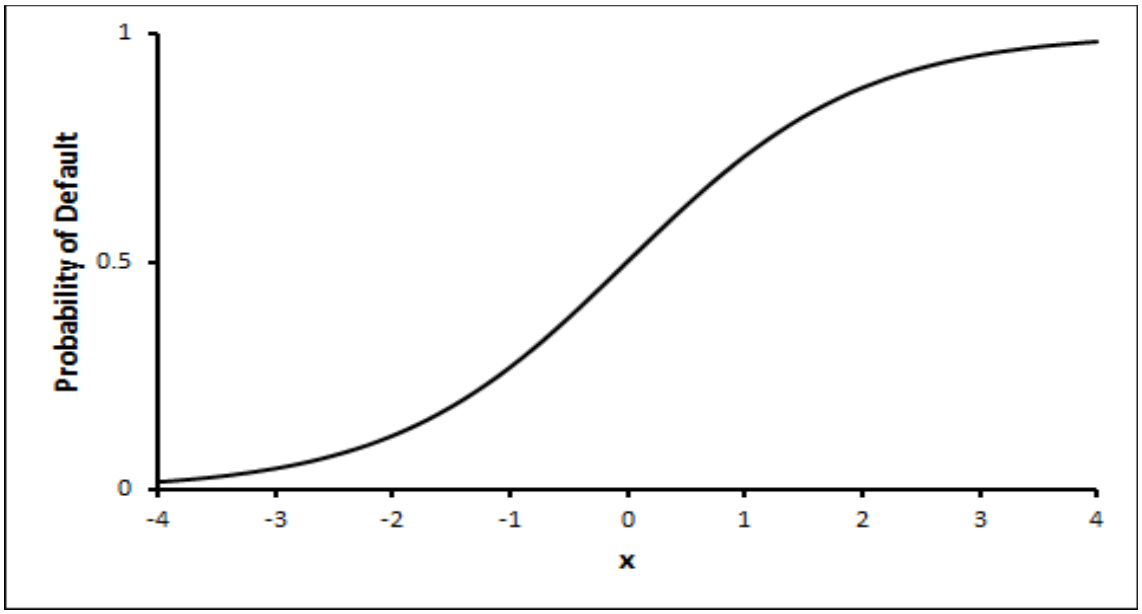


Figure 2.5: An example of logistic function.

binary dependent variable. The predicted values of a logistic regression model are probabilities that predicts the probability of a certain outcome, and are therefore restricted to be contained in [0,1]. Logistic regression measures the relationship between a binary dependent variable and one or more independent variables by estimating probabilities using a logistic function as presented as the S-shape curve in Figure 2.5. In our case, the dependent variable Y equals 1 for bad payers and equals 0 for good payers. The curve as presented in Figure 2.5 can be defined by the equation

$$p(\mathbf{x}) = \frac{\exp(\beta\mathbf{x})}{1+\exp(\beta\mathbf{x})},$$

where $p(\mathbf{x})$ is the proportion of borrowers that are bad payers. The probabilities that $Y = 1$ and $Y = 0$ can be expressed as in Eqs. (2.56) and (2.57):

$$P(Y = 1|\mathbf{X}) = p(\mathbf{x}) = \frac{\exp(\beta\mathbf{x})}{1+\exp(\beta\mathbf{x})} \tag{2.56}$$

$$P(Y = 0|\mathbf{X}) = 1 - p(\mathbf{x}) = \frac{1}{1+\exp(\beta\mathbf{x})}. \tag{2.57}$$

Here $\exp(\cdot)$ denotes the exponential function, β represents a vector of parameters and $\mathbf{x}$ represents the attributes of each borrower. Rearranging Eq. (2.56) gives:

$$\frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \exp(\beta\mathbf{x})$$

Taking the natural logarithm on both sides, it becomes

$$G(\mathbf{x}) = \ln\left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right) = \beta\mathbf{x}. \tag{2.58}$$

This transformation, $G(\mathbf{x})$ is called the logit transformation. Notice that $G(\mathbf{x})$ has many of the desirable properties of a linear regression model. It is linear in its parameters, and $G(\mathbf{x})$ can have values ranging from $-\infty$ to ∞, depending on the values of $\mathbf{x}$. Notice that logistic regression is trying to fit $\ln\left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right)$ by a linear combination of the attributes. Just as when using discriminant analysis, the logistic regression model will be developed using historical data with known repayment performance. In most ordinary regression, least-squares approach will be applied to calculate the coefficients β, however, for logistic regression, there is no mathematical solution that will produce explicit expressions for least square estimates of the parameters. The approach that will apply to estimate the parameters is called maximum likelihood. A method that yields values for the unknown parameters that maximize the probability of obtaining the observed set of data. A likelihood function which express the probability of the observed data as a function of the unknown parameters will be constructed, then the maximum likelihood estimators of these parameters will be chosen so as to maximize the likelihood function. Again, $p(\mathbf{x})$ gives the conditional probability $P(Y = 1|X)$ and $1 - p(\mathbf{x})$ gives the conditional probability $P(Y = 0|X)$. For an observation $(y_i, \mathbf{x}_i)$ where $y_i = 1$, the contribution to the likelihood function is $p(\mathbf{x})$ and where $y_i = 0$, the contribution to the likelihood function is $1 - p(\mathbf{x})$. Therefore, for the observation $(y_i, \mathbf{x}_i)$, we know that

$$P(Y = y_i) = p(\mathbf{x}_i)^{y_i}(1 - p(\mathbf{x}_i))^{1-y_i}$$

With the assumption that all the observations are independent, the likelihood function can be expressed as

$$l(\beta) = \prod_{i=1}^{n} p(\mathbf{x}_i)^{y_i}(1 - p(\mathbf{x}_i))^{1-y_i}$$

The log likelihood function can be defined as:

$$L(\beta) = \ln(l(\beta)) = \sum_{i=1}^{n} [y_i \ln(p(\mathbf{x}_i)) + (1 - y_i)\ln(1 - p(\mathbf{x}_i))] \tag{2.59}$$

Notice that the above log likelihood function is nonlinear with the parameters β, therefore numerical methods must be use to obtain their solutions. We will not go into the details on estimating the parameters for logistic regression models. However, we suggest the use of the statistical software R, to generate logistic regression models by using the function glm() and

set family equals to binomial. For further information on using R to develop logistic regression models, one can refer to Hothorn and Everitt [17].

Note that logistic regression is more of a data-driven approach, for which a given dataset will be fitted to the logistic function as presented in Figure 2.5. Then a model will be developed regardless of the distribution of the underlying data. The application of maximum likelihood to estimate the parameters is not direct, instead, numerical methods must be applied to solve for a solution for the parameters. In this thesis, we model the interaction between the banks and borrowers, particularly examining the effect of bad borrowers' added lies onto their reported attributes. We study our research by assuming the distribution of the characteristics of borrowers are known. We then imagine the effect of bad borrowers who add lies and the resulting effect by the banks to eliminate these lies by altering the distribution of the characteristics. With full knowledge about the distribution of the data and the parameters used in the model, we choose discriminant analysis, a model-driven approach, as the modeling method that will be used throughout this thesis. Discriminant analysis provides a clear picture on directly classifying borrowers into two groups of good or bad payers. We can fully understand the change of data and model when lies are included. Therefore, we can focus on analyzing the effect of borrowers lies towards the profitability of banks.

2.3 Conclusions

Many different statistical techniques are applied in the context of credit scoring. This chapter specifically detailed the use of discriminant analysis, a method that will be applied throughout succeeding chapters of this thesis. We analyzed the discriminatory power of discriminant analysis by assuming the characteristics of borrowers follows univariate and bivariate normal distribution. Discriminant analysis can be extended to higher dimensional data. Indeed, as long as the probability density functions of the characteristics of good and bad payers are known, discriminant analysis can be applied whatever the dimension of the attributes set. Later in Chapter 5 we apply discriminant analysis ,by assuming the underlying distribution follows right triangular distribution, to classify good and bad borrowers. In addition, we provided an overview of logistic regression analysis, another well-known method used in the application of credit scoring. The development of more advanced and accurate statistical models to classify borrowers present challenging opportunities for future statisticians.

Bibliography

[1] T. S. Lee, C. C. Chiu, C. J. Lu and I. F. Chen, Credit scoring using the hybrid neural discriminant technique, *Expert Systems with applications*, **23**(3)(2002), pp. 245-254.

[2] C. Bravo, L. C. Thomas and R. Weber, Improving credit scoring by differentiating defaulter behavior, *Journal of the Operational Research Society*, **66**(5)(2014), pp.771-781.

[3] L. Thomas, D. B. Edelman and J. N. Crook, *Credit Scoring and Its Applications*, SIAM, Philadelphia, USA, 2002.

[4] E. I. Altman, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *The Journal of Finance*, **23**(4)(1968), pp. 589-609.

[5] L. C. Thomas, A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers, *International Journal of Forecasting*, **16**(2)(2000), pp. 149-172.

[6] A. Steenackers and M. J. Goovaerts, A credit scoring model for personal loans, *Insurance: Mathematics and Economics*, **8**(1)(1989), pp. 31-34.

[7] R. B. Avery, P. S. Calem and G. B. Canner, Consumer credit scoring: do situational circumstances matter?, *Journal of Banking & Finance*, **28**(4)(2004), pp. 835-856.

[8] Basel Committee on Banking Supervision, June 2004. Basel II: International Convergence of Capital Measurement and Capital Standards. Bank for International Settlements.

[9] D. J. Hand and W. E. Henley, Statistical classification methods in consumer credit scoring: a review, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **160**(3)(1997), pp. 523-541.

[10] D. J. Hand, Reject inference in credit operations, *Credit risk modeling: Design and application* (1998), pp. 181-190.

[11] B. Anderson, S. Haller and N. Siddiqi, *Reject Inference Techniques Implemented In Credit Scoring for SAS Enterprise Miner*, SAS Working paper, 305, `http://support.`

sas.com/resources/papers/proceedings09/305-2009.pdf, dostep: 2014.03. 23, 2009.

[12] B. Ripley, B. Venables, D. M. Bates, K. Hornik, A. Gebhardt, D. Firth and M. B. Ripley, Package "MASS". CRAN Repository. http://cran.r-project.org/web/packages/MASS/MASS.pdf(2013).

[13] S. G. Johnson and B. Narasimhan, Adaptive multivariate integration over hypercubes. http://cran.r-project.org/web/packages/cubature/cubature.pdf, February 2013.

[14] J. Berntsen, T. O. Espelid and A. Genz, An adaptive algorithm for the approximate calculation of multiple integrals. *ACM Transactions on Mathematical Software (TOMS)*, **17**(4)(1991), pp. 437-451.

[15] J. C. Wiginton, A note on the comparison of logit and discriminant models of consumer credit behavior, *Journal of Financial and Quantitative Analysis*, **15**(03)(1980), pp.757-770.

[16] M. Bensic, N. Sarlija and M. Zekic-Susac, Modelling small-business credit scoring by using logistic regression, neural networks and decision trees, *Intelligent Systems in Accounting, Finance and Management*, **13**(3)(2005), pp. 133-150.

[17] T. Hothorn and B. S. Everitt, *A handbook of statistical analyses using R*, CRC press, 2014.

[18] B. Efron, The efficiency of logistic regression compared to normal discriminant analysis, *Journal of the American Statistical Association*, **70**(352)(1975), pp. 892-898.

[19] S. J. Press and S. Wilson, Choosing between logistic regression and discriminant analysis, *Journal of the American Statistical Association*, **73**(364)(1978), pp. 699-705.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014, https://www.R-project.org/.

Chapter 3

More Attributes May Lead to More Inaccurate Credit Scoring

3.1 Introduction

Existing credit scoring models are built using as much historical data as possible from past borrowers with known repayment performance. Analysts believe that using more attributes to build credit scoring models will always increase model accuracy. However, Oates and Jensen [1] show that increasing the amount of data used to build models may result in a more complex model with no significant increase in accuracy. We will show that contrary to the model, if some borrowers respond untruthfully to some questions, using higher dimensional data may even reduce the model's predictive power compared against using a dataset with lower dimensions. Using more data to build the model will increase the associated accumulated error and results in an overestimated model with low accuracy. The proposed issue will be studied using simulated data and discriminant analysis based on the credit scoring context. In fact, knowing the optimal amount of data that is required to produce accurate predictions, the questionnaire used to collect information from borrowers can be shortened, which reduces the time cost associated with collecting, checking and processing the data. This research can enhance the efficiency of the granting decisions and help the lending financial institutions to maximize profit in their loan activities.

The succeeding sections are organized as follows. Section 3.2 presents the method and model used for this problem. In Section 3.3, the data used to analyze the problem was simulated. In Section 3.4, we applied classification techniques to Section 3.3's data to support the results and provide an explanation on how to visualize the results. Conclusions are drawn in Section 3.5.

3.2 Method and Model

This chapter uses discriminant analysis as presented in Section 2.1 of this thesis. Three different models have been considered. The first model considers only one continuous normally distributed characteristic X. We presented the model as equation (2.7) of the previous chapter, showing that the set of good payers must have characteristics x satisfying the inequality below.

Model 1

$$A_G = \left\{ x \mid \frac{(x-\mu_G)^2}{\sigma_G^2} - \frac{(x-\mu_B)^2}{\sigma_B^2} \leq -2\log\left(\frac{Lp_B\sigma_G}{Qp_G\sigma_B}\right) \right\}.$$

The second model considers two continuous normally distributed characteristics X_1 and X_2. We built our second model based on equation (2.14). In order to simplify our calculations, we set $\rho = 0$. The corresponding density function for good and bad payers are:

$$f(x_{1G}, x_{2G}) = \frac{1}{2\pi\sigma_{1G}\sigma_{2G}} \exp\left\{\frac{-(x_1-\mu_{1G})^2}{2\sigma_{1G}^2}\right\} \exp\left\{\frac{-(x_1-\mu_{2G})^2}{2\sigma_{2G}^2}\right\}$$

and

$$f(x_{1B}, x_{2B}) = \frac{1}{2\pi\sigma_{1B}\sigma_{2B}} \exp\left\{\frac{-(x_1-\mu_{1B})^2}{2\sigma_{1B}^2}\right\} \exp\left\{\frac{-(x_1-\mu_{2B})^2}{2\sigma_{2B}^2}\right\}.$$

Applying set condition (2.5), the right hand side of the inequality becomes

$$\begin{aligned}\frac{f(x|G)}{f(x|B)} &= \frac{\frac{1}{2\pi\sigma_{1G}\sigma_{2G}} \exp\left\{\frac{-(x_1-\mu_{1G})^2}{2\sigma_{1G}^2}\right\} \exp\left\{\frac{-(x_1-\mu_{2G})^2}{2\sigma_{2G}^2}\right\}}{\frac{1}{2\pi\sigma_{1B}\sigma_{2B}} \exp\left\{\frac{-(x_1-\mu_{1B})^2}{2\sigma_{1B}^2}\right\} \exp\left\{\frac{-(x_1-\mu_{2B})^2}{2\sigma_{2B}^2}\right\}} \\ &\geq \frac{Lp_B}{Qp_G}.\end{aligned}$$

From the above, we conclude our second model showing the set of good payers must have characteristics x_1 and x_2 satisfying the inequality below.

Model 2

$$A_G = \left\{ \mathbf{x} \middle| \frac{\frac{1}{2\pi\sigma_{1G}\sigma_{2G}} \exp\left\{\frac{-(x_1-\mu_{1G})^2}{2\sigma_{1G}^2}\right\} \exp\left\{\frac{-(x_1-\mu_{2G})^2}{2\sigma_{2G}^2}\right\}}{\frac{1}{2\pi\sigma_{1B}\sigma_{2B}} \exp\left\{\frac{-(x_1-\mu_{1B})^2}{2\sigma_{1B}^2}\right\} \exp\left\{\frac{-(x_1-\mu_{2B})^2}{2\sigma_{2B}^2}\right\}} \overset{\text{set}}{\geq} \frac{Lp_B}{Qp_G} \right\}. \tag{3.1}$$

3.3 Simulation of Data

We use simulated data to illustrate our proposed issue, that of prospective borrowers lying about one or more of their attributes. We apply the technique presented in Appendix A, with certain modifications to simulate our dataset. In particular, we want to simplify our calculations

and results by assuming X_1 and X_2 are independent, hence, $\rho = 0$. The procedure for simulating the dataset is described below:

Step 1. $X_1 = \mu_1 + \sigma_1 Z_1$, where $Z_1 \sim N(0,1)$

Step 2. $X_2 = \mu_2 + \sigma_2 Z_2$, where $Z_2 \sim N(0,1)$ and $Z_1 \perp Z_2$

We make the (strong) assumption that only bad payers will lie about their information and can only alter their characteristics through X_{2B}, with the objective of making themselves indistinguishable from the good payers. To model this we added noise to X_{2B} only. We believe that not all bad payers choose to lie about their characteristics; and modeled this by introducing another Bernoulli random variable *Noi*, which takes the value one when a particular bad payer lies, otherwise taking the value zero. We equate the probability that a bad payer will lie to pN (or *Noi* takes the value 1). We further assume that all bad payers who intend to lie do so by adding a fixed constant amount A to their attribute. For example, a liar might say they earned 10K more than they actually did. As a result,

$$X_{2Bnew} = X_{2B} + Noi \times A \tag{3.2}$$

μ_{1G}	σ_{1G}	μ_{1B}	σ_{1B}	μ_{2G}	σ_{2G}	μ_{2B}	σ_{2B}	L	Q	p	pN	A
8	1	4	1	8	1	6	1	5	2	0.5	0.9	$\frac{2}{0.9}$

Table 3.1: Parameters used in the case study described in Section 3.4

3.4 Results and Analysis

Notice that model 1 and model 2 can be applied to do prediction if $f(\mathbf{X}|G)$ and $f(\mathbf{X}|B)$ are normally distributed. However, our data is not in fact normal, rather a mixture of normals. So we see if our data nonetheless appear normal enough that a busy credit quant might model it as normal. A proportion of lies is added to X_{2B} as described in Section 3.3. To ensure the plausibility of an incorrect normal assumption for X_{2Bnew}, we generated a histogram and a Q-Q plot for X_{2Bnew}, and performed Anderson-Darling (Stephens, 1986 [2]; Thode, 2002, Sec. 5.1.4 [3]) and Cramer-von Mises (Stephens, 1986 [2]; Thode, 2002, Sec. 5.1.3 [3]) normality tests to show that the possibility that X_{2Bnew} still follows a normal distribution cannot be rejected. The

Q-Q plot in Figure 3.1 and the two test results in Table 3.2 ensured that X_{2Bnew} is sufficiently close to a normal distribution to fool many observers.

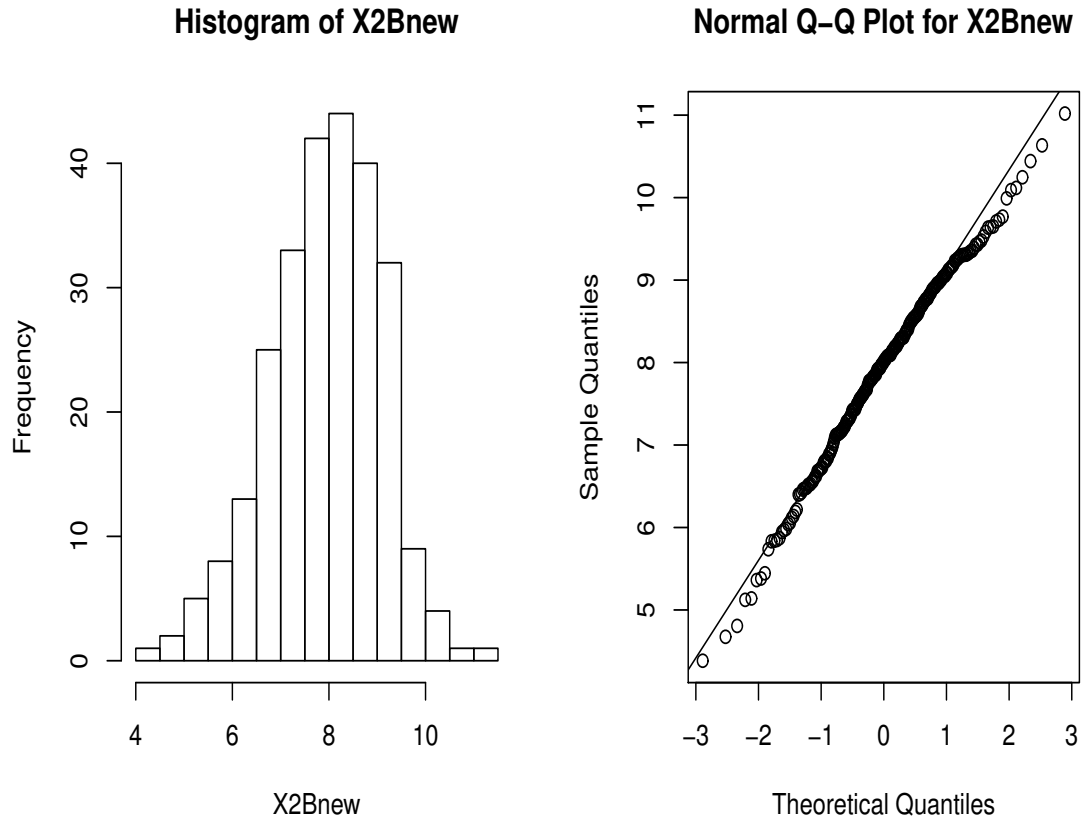


Figure 3.1: Data was simulated using the method described in Section 3.3 and the parameters in Table 3.1. Constant amount of lies was added to X_{2B} according to equation (3.2) to generate X_{2Bnew}. The plot of histogram and Q-Q plot provide evidence that it is possible to misinterpret X_{2Bnew} as normally distributed.

Using the parameters of Table 3.1, we simulated 100 datasets each with 500 borrowers and reran the program 100 times. Therefore, the total number of simulations is 10,000. We use model 1 and model 2 to find the accuracy of the prediction and the profit earned. Note that, in an attempt to keep historical computation representative, classes may have relatively few members. Let GG be the number of predicted good payers that are truly good payers, BB be the number of predicted bad payers that are truly bad payers, GB be the number of predicted good payers that are truly bad payers and BG be the number of predicted bad payers that are truly good payers. The average values of the output from the program are stated in Table 3.3. The accuracy of classification was defined as $\frac{(GG+BB)}{\text{Total number of borrowers}}$. The profit earned was defined as $Q \times GG - L \times GB$, where Q is the expected profit that the lender will earn if the

Anderson-Darling Test	Cramer-von Mises Test
P-value = 0.101	P-value = 0.134

Table 3.2: Both normality tests provide p-value bigger than 0.05. There is not enough evidence to reject the null hypothesis and conclude that X_{2Bnew} is normally distributed.

borrower is a good payer, and L is the expected loss that the lender will incur if the borrower is a bad payer. We computed the average accuracy and profit of the 100 datasets and also found the standard error of those values. From the results shown in Table 3.3, there is evidence that using only one characteristic to do prediction is on average 1.74% better and earns $57 more than using two characteristics.

In fact X_{2Bnew} is not normally distributed; we can find the true distribution of X_{2Bnew} and insert that into discriminant analysis to obtain the correct results. We know that $\rho = 0$, X_{1B} and X_{2Bnew} are independent and Noi is independent of X_{1B} , therefore

$$\frac{f(\mathbf{x}|G)}{f(\mathbf{x}|B)} = \frac{f(x_{1G}, x_{2G})}{f(x_{1B}, x_{2B})} = \frac{f(x_{1G})f(x_{2G})}{f(x_{1B})f(x_{2B})}.$$

Through the way we generated X_{2Bnew}, we can deduce that

$$\begin{aligned} f(x_{2Bnew}) &= f(x_{2Bnew}|Noi = 0)P(Noi = 0) + f(x_{2Bnew}|Noi = 1)P(Noi = 1) \\ &= (1 - pN)\frac{1}{\sqrt{2\pi\sigma_{2B}^2}} \exp\left\{\frac{-(x_2 - \mu_{2B})^2}{2\sigma_{2B}^2}\right\} + pN\frac{1}{\sqrt{2\pi\sigma_{2B}^2}} \exp\left\{\frac{-(x_2 - \mu_{2Bnew})^2}{2\sigma_{2B}^2}\right\}, \end{aligned}$$

where μ_{2Bnew} equals $\mu_{2B} + A$.

Applying inequality (2.5), we found our third model presenting the correct set condition and giving the set of good payers as below.

Model 3

$$A_G = \left\{ \mathbf{x} \middle| \frac{\frac{1}{\sigma_{1G}\sigma_{2G}} \exp\left\{\frac{-(x_1-\mu_{1G})^2}{2\sigma_{1G}^2}\right\} \exp\left\{\frac{-(x_2-\mu_{2G})^2}{2\sigma_{2G}^2}\right\}}{\frac{1}{\sigma_{1B}\sigma_{2B}} \exp\left\{\frac{-(x_1-\mu_{1B})^2}{2\sigma_{1B}^2}\right\}\left[(1-pN)\exp\left\{\frac{-(x_2-\mu_{2B})^2}{2\sigma_{2B}^2}\right\} + pN\exp\left\{\frac{-(x_2-\mu_{2Bnew})^2}{2\sigma_{2B}^2}\right\}\right]} \geq \frac{Lp_B}{Qp_G} \right\}. \tag{3.3}$$

Using the above inequality, the correct accuracy of using two characteristics is 97.46% and it earns $469.19, which is 0.05% better and earns $0.66 more than using only one characteristic. This coincides with the well accepted intuition that using more information produces better performance. Note that when comparing model 1 and model 3, the simple one-attribute classi-

fier did not show significant degradation in the accuracy: model 3 gives an accuracy of 97.46% and profit of $469.19, while the significantly simpler model 1 gives an accuracy of 97.41% and profit of $468.52. According to those figures, using two variables only shows minor improvements on the accuracy of prediction and profit earned, compared against using only one variable. Even here, it is likely wise to use only one variable to lower the cost and speed the analysis process.

	Model 1	**Model 2**	**Model 3** **(Corrected Model)**
Attribute	X_1 only	X_1 and X_2	X_1 and X_2
BB	244.72 ± 2.45	231.42 ± 2.40	244.79 ± 2.40
BG	9.70 ± 0.67	5.08 ± 0.52	9.55 ± 0.65
GB	3.24 ± 0.38	16.53 ± 0.74	3.16 ± 0.36
GG	242.35 ± 2.51	246.97 ± 2.39	242.50 ± 2.47
Accuracy	97.41% ± 0.15%	95.68% ± 0.19%	97.46% ± 0.16%
Profit	468.52 ± 5.11	411.26 ± 6.37	469.19 ± 5.30

Table 3.3: Using the parameters in Table 3.1 and the simulation method described in Section 3.3. The results are generated using model 1, 2 and 3 are presented above. The first values shows the average value of the 100 datasets, the value after the "±" sign shows the standard error of the average value.

In order to have a clear understanding of the results, we try to visualize the performance of the cutoff line on classifying the good and bad payers. We evaluated the cutoff lines for model 2 and model 3 numerically. Inserting the initial conditions from Table 3.1 into equation (3.1) of model 2, the set condition becomes

$$A_G = \left\{ \mathbf{x} \mid x_1 \geq \frac{1}{4}\left[38 + \log\left(\frac{5}{2}\right) - 2x_2 \right] \right\},$$

which leads to the cutoff line

$$x_1 = \frac{1}{4}\left[38 + \log\left(\frac{5}{2}\right) - 2x_2 \right]. \tag{3.4}$$

Putting the initial conditions into equation (3.3) of model 3, the set condition becomes

$$A_G = \left\{ \mathbf{x} \mid x_1 \geq \frac{1}{4}\left[24 + \log\left(\frac{\frac{5}{2}\left[0.1\exp\left\{ \frac{-(x_2-6)^2}{2} \right\} + 0.9\exp\left\{ \frac{-(x_2-6-\frac{2}{0.9})^2}{2} \right\} \right]}{\exp\left\{ \frac{-(x_2-8)^2}{2} \right\}} \right) \right] \right\},$$

and that leads to the cutoff line

$$x_1 = \frac{1}{4}\left[24 + \log\left(\frac{\frac{5}{2}\left[0.1\exp\left\{\frac{-(x_2-6)^2}{2}\right\} + 0.9\exp\left\{\frac{-(x_2-6-\frac{2}{0.9})^2}{2}\right\}\right]}{\exp\left\{\frac{-(x_2-8)^2}{2}\right\}}\right)\right]. \tag{3.5}$$

We divide all the borrowers into two groups of good and bad payers, and plot the contours of the two characteristics X_1 and X_2 of the two groups, together with the cutoff lines of model 2 and 3 in Figure 3.2. The right hand side of the cutoff lines represent the region where the model will classify the payers as good, while the left hand side of the cutoff lines represent the region where the model will classify the payers as bad. From Table 3.3, it shows that there are more borrowers who are predicted to be bad while actually being good in model 3 than in model 2, and there are more borrowers who are predicted to be good while truly being bad in model 2 than in model 3. Those results are reflected in the region between the solid line and the dotted line in Figure 3.2. There are more bad payers classified as good using the solid line, and there are more good payers classified as bad using the dotted line.

3.5 Conclusions

Many researchers believe that using more attributes to develop classification model will always increase model accuracy. In this chapter, we showed that using fewer attributes to predict good or bad borrowers is more reasonable and produce more accurate results due to the fact that people misused prediction models in the classification process. More effort should be expended on examining the data to detect fraud and on whether the prediction models are statistically valid. In Chapter 4, we will further study the effect on banks' profitability of bad borrowers purposely adding lies onto their reported attributes in order to obtain. The correct usage of credit scoring models can effectively and efficiently fulfill the needs of processing loan requests and can maximize the profit from the lending of money.

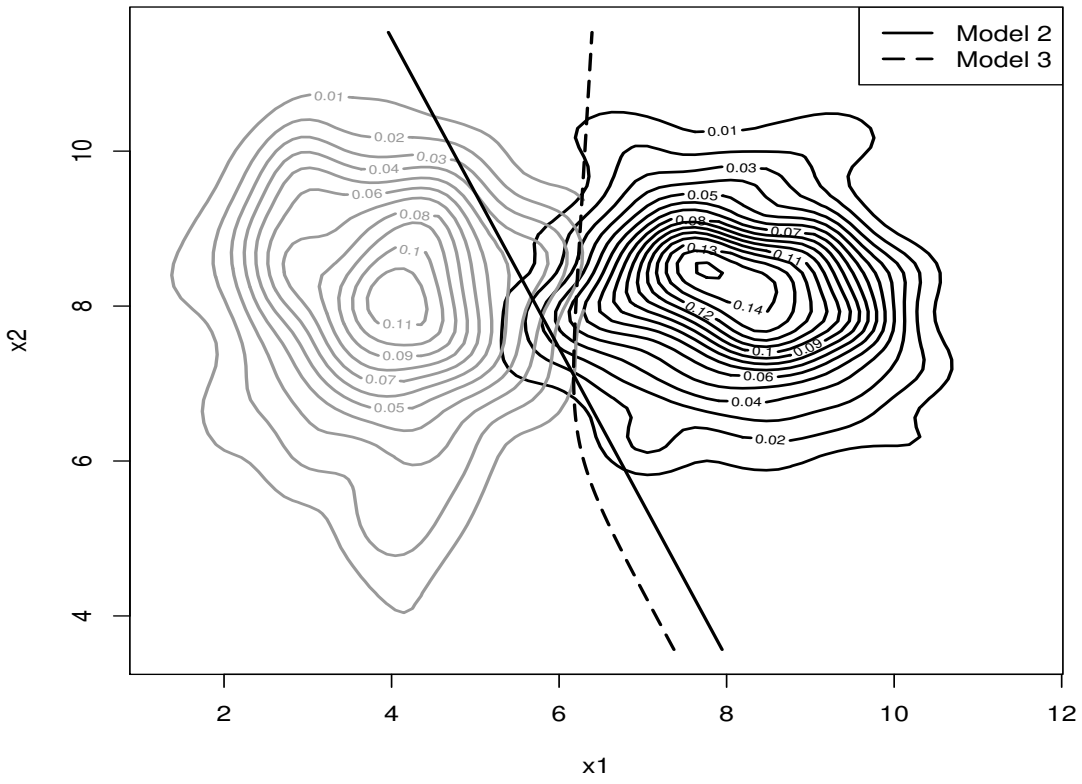


Figure 3.2: The gray contour on the left displays the characteristic values of the bad payers, while the black contour on the right displays the characteristic values of the good payers. The solid line is the cutoff line plotted using equation (3.4), while the dotted line is the cutoff line plotted using equation (3.5).

Bibliography

[1] T. Oates and D. Jensen, Large Datasets Lead to Overly Complex Models: An Explanation and a Solution, *KDD*, (1998), pp. 294-298.

[2] M. A. Stephens, Tests based on EDF statistics, *Goodness-of-fit Techniques*, **68**(1986), pp. 97-193.

[3] H. C. Thode, *Testing for normality* (Vol. 164). CRC press, 2002.

Chapter 4

How Much Effort Should Be Spent to Detect Fraudulent Applications When Engaged in Classifier-Based Lending?

This chapter is based on the paper [1] that we published in the Intelligent Data Analysis Journal. It contains an additional set of calculations and some reordering of ideas for clearer exposition. The details on the method and model used in this research can be found in Chapter 2 of this thesis and will not be explicitly repeated again in this Chapter. The succeeding sections are organized as follows. Section 4.1 gives a brief introduction of this chapter. Section 4.2 presents the method and model used for this problem. In Section 4.3, the data used to analyze the problem was simulated. In Sections 4.4 and 4.5, we considered two different choices on the amount of loan granted to borrowers, we discuss and compare the results from the banks' prospective. Section 4.6 presents and analyzes the results from borrowers' prospective. Conclusions are drawn in Section 4.7.

4.1 Introduction

In recent years the use of machine learning or classifiers on credit applications has become a killer app for retail and small commercial lending. Many classifier algorithms are used for credit scoring, as reviewed in Baesens *et al.* [1] and in Lessmann *et al.* [5]. However, as always, the classifier can only do a good job if the attributes it employs to make its classifications are at least nearly correct. As discussed in Dejaeger *et al.* [4], the accuracy and reliability of data is of high importance, and in the cases of insufficient data quality, quantitative analysis may fail to provide the desired results, even leading to incorrect decisions being drawn. This is particularly

a problem for the use of classifiers to grant loans, where prospective borrowers may have a strong motivation to make a fraudulent application by falsifying one or more of the attributes they report on their application forms. It is generally expensive to check credit applications for lies. In Chapter 1, we introduced the concept of bust-out-fraud, where borrowers purposely build up and maintain a good credit history, they will then use all the available credit in one time to borrow a large amount of loan, and then abandon the account. We consider this in some sense to be like lying about the attribute since behavior is modified to get a desired score rather than for its own intrinsic rewards.

In this chapter, we will look into the problem of having prospective borrowers fraudulently falsify one or more of the attributes they report on their application form. Applicants learn about the characteristics that are used to build credit scoring models, and may alter the answers on their application form to improve their chance of loan approval. Few automated credit scoring models have considered falsified information from borrowers. We will show that sometimes it is profitable for financial institutions to spend money and effort to identify dishonest customers. We will also find the optimal effort that banks should spend on identifying these liars. Furthermore, we will show that it is possible for liars to eventually adjust their lies to escape from credit checks. This research uses simulated data and discriminant analysis as presented in Section 2.1 of this thesis to study the proposed issue.

For simplicity, we will restrict our studies on lies that borrowers placed on one numerical attribute used in the classifier. Indeed, we assume the loan application involves only one attribute. In our case, a larger attribute value is better for getting a loan. Not all attributes are like this (for instance applicants' total existing debt), in which case multiply the attribute by -1 to get our convention. We make an assumption that the size of lie is the difference between the correct value and the reported value. In real situations, customers do not know whether they will be identified as a good or a bad payer. On the other hand, lenders do not know the future performance of borrowers. Therefore, there are four different cases:

1. Good borrowers who are honest on their loan application form and will always repay their loans.

2. Good borrowers who lie on their loan application form but always repay their loans.

3. Bad borrowers who are honest on their loan application form but do not repay their loans.

4. Bad borrowers who lie on their loan application form and do not repay their loans.

Since we build our model using simulated data meant to model historical data, we have precise information about whether a given borrower repaid her loans (good) or failed to repay (bad). In addition, we assume we know whether the borrowers lied about their characteristics.

In particular, we assume all good borrowers are honest and all borrowers who lied defaulted. In other words, we assumed our dataset contains borrowers from cases 1, 3 and 4.

It is straightforward to see that the effort which should be devoted to checking lies depends on the size of the lies, the importance of the attribute lied about to the classifier, the cost of reducing the magnitude of lies through better checking, and the proportion of bad liars in the population. In this chapter, we present a proof of concept model which incorporates all these elements. A simulation study applied to this model answers the questions: 1) For a given lie strength and lie detection cost, can it be worthwhile to expand effort to decrease lies? If yes, 2) What level of lie elimination effort is optimal? In order to answer 1) and 2) the related question 3) What is the optimal level of lies for an unscrupulous borrower? is also discussed. In particular, this research shows that when considering the optimal effort required to reduce fraud, the financial institution should consider the optimal degree to which borrowers lie.

Misreporting, that is, the impact of manipulation in credit risk, is a topic that has recently attracted attention from the research community. Most studies focus on corporate lending and securitized loans, such as Griffin and Maturana [8] and Graham and Qiu [7]. On unsecuritized consumer loans, the work of Garmaise [14] centers on the increase in delinquency that arises from misreporting, rather than the incentives that are in place when requesting loans.

Other studies have focused more on the reasons that drive default. The work of Bravo *et al.* [4] suggested that defaulters behave strategically, and divides defaulters into a group that cannot repay due to financial problems and another group who perhaps never intended to repay. Guiso *et al.* [11] also found that borrower default is not rational; rather it is caused by economic and moral factors. In fact, borrowers can be strategic before requesting loans. Our work builds from these themes.

Credit scoring models have been widely used in different lending industries. Borrowers can learn about the characteristics that lending institutions will use to make granting decisions. Borrowers may also determine how to alter their claimed characteristics so as to increase their chance of loan approval. Spending money on detecting and eliminating the lies that borrowers made in the data will increase the cost of implementing credit scoring models. The purpose of this chapter is to show that, in the context of a detection cost model, it may be profitable to reduce the magnitude of lies in the dataset. In fact, spending effort to reduce the magnitude of lies in the dataset which was used to build credit scoring models can increase the accuracy of prediction and lower the risk of lending money. However, verifying the accuracy of the data is expensive, and the cost to improve the data should be considered. In addition, we will show that liars will ultimately learn how to adjust their lies in order to escape from credit checks. Therefore, regular updates of the scoring model are necessary to maintain accurate credit decisions.

4.2 Method and Model

This chapter uses discriminant analysis as presented in Section 2.1, and considers $\mathbf{X}$ to be a single continuous attribute drawn from a normal distribution, where $X_G \sim N(\mu_G, \sigma_G)$ and $X_B \sim N(\mu_B, \sigma_B)$ same as in Section 2.1.1. Again, the probability density function for good payers is

$$f(x|G) = \frac{1}{\sqrt{2\pi}\sigma_G} \exp\left\{\frac{-(x-\mu_G)^2}{2\sigma_G^2}\right\}. \tag{4.1}$$

We assume that only bad payers will lie about their attribute and can alter their characteristic through X_B. We assume that good payers do not lie, and so added noise to X_B only. Not all bad payers choose to lie about their characteristics; we introduced another Bernoulli random variable *Noi*, which takes the value one when a particular bad payer lies, otherwise taking the value zero. The probability that a bad payer will lie is P_N. We further assume that all bad payers who intended to lie do so by adding a constant A to the correct attribute value. In addition, we assume a parameter η which represents the effort spent on reducing lies in characteristic X_B, to reduce the magnitude of lies to $A(1-\eta)$. Note that in our case η will be a value between 0 to 1. For example, a liar who earned \$50,000 annually might say that he earned \$60,000. In this case, *Noi* equals 1, since the borrower lied and A would be \$10,000. If the lender did not put any effort to check the attribute value of the borrower ($\eta = 0$), then the reported attribute will equal \$60,000, if the lender put some effort ($\eta = 0.5$) on checking the attribute, then the reported value will be \$55,000, closer to the correct value. If $\eta = 1$, that refers to the case where the lender carefully checks the attribute values reported by the borrowers and successfully changed the reported value to its correct value. In real situations, it is extremely difficult for lenders to exclude all the lies that borrowers make. However, the chances of having all the lies excluded for a particular borrower is not impossible, and our model captures that. Furthermore, the incorrect value of the attribute may not be the cause of an intended lie. For instance, it is difficult to determine the value of a house: different appraisers might assign a slightly different value for the same house. Therefore, even if the lender found some discrepancies between the reported value and the true value of the borrower's attribute, as long as the lender truly thinks the borrower has a high chance of repaying, the loan should be approved. Note that we simplified our chapter by assuming all discrepancies from the true value of an attribute are due to a lie. As a result,

$$X_{Bnew} = X_B + Noi \times A(1-\eta), \quad \text{where} \quad Noi \sim \text{Bern}(P_N).$$

[Note that X_B is independent of *Noi*.]

Notice that, after including the lies of the bad payers and the effort spent on eliminating

lies, the cumulative distribution function

$$\begin{aligned}
F_{X_{Bnew}}(x) &= P(X_{Bnew} \leq x) \\
&= P(X_{Bnew} \leq x|Noi = 0)P(Noi = 0) + P(X_{Bnew} \leq x|Noi = 1)P(Noi = 1) \\
&= P(X_B \leq x)P(Noi = 0) + P(X_B + A(1-\eta) \leq x)P(Noi = 1) \\
&= P(X_B \leq x)P(Noi = 0) + P(X_B \leq x - A(1-\eta))P(Noi = 1) \\
&= (1 - P_N) \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma_B^2}} \exp\left\{\frac{-(z-\mu_B)^2}{2\sigma_B^2}\right\} dz \\
&+ P_N \int_{-\infty}^{x-A(1-\eta)} \frac{1}{\sqrt{2\pi\sigma_B^2}} \exp\left\{\frac{-(z-\mu_B)^2}{2\sigma_B^2}\right\} dz.
\end{aligned}$$

Hence, by taking the derivative with respect to x the probability density function of X_{Bnew} becomes

$$f(x|Bnew) = (1 - P_N)\frac{1}{\sqrt{2\pi\sigma_B^2}} \exp\left\{\frac{-(x-\mu_B)^2}{2\sigma_B^2}\right\} + P_N \frac{1}{\sqrt{2\pi\sigma_B^2}} \exp\left\{\frac{-(x-\mu_{Bnew})^2}{2\sigma_B^2}\right\}, \quad (4.2)$$

where $\mu_{Bnew} = \mu_B + A(1-\eta)$.

For simplicity we assume no recovery in the event of default, so the expected loss L is the loan granted to the borrower. In general, we denote L_i as loan amount requested by borrower i. Therefore, the amount that borrower i must repay is $Q_i = L_i \times I$, where I is the interest rate. Applying the set inequality in Eq. (2.5) of Section 2.1 and using the above information, we can conclude our classification model giving the set of attributes reported by good payers is

$$A_G = \left\{ x_i \middle| \frac{L_i p_B}{Q_i p_G} \leq \frac{f(x_i|G)}{f(x_i|Bnew)} \right\}, \quad (4.3)$$

where $f(x_i|G)$ and $f(x_i|Bnew)$ are stated in Eqs. (4.1) and (4.2) respectively. We consider two different choices on the loan amount granted to borrowers and will explained in detail in Section 4.3. Substituting $Q_i = L_i \times I$ into Eq. (4.3) resulted in

$$A_G = \left\{ x_i \middle| \frac{p_B}{I p_G} \leq \frac{f(x_i|G)}{f(x_i|Bnew)} \right\}. \quad (4.4)$$

Note that the above classification model is independent of the loan amount granted L_i.

4.3 Simulation of Data

We use simulated data to illustrate our proposed issue, that of prospective borrowers lying about one of their attributes, and the effect of the corresponding effort spent on eliminating lies. We apply the technique presented in Appendix A, with certain modifications to simulate our dataset. In particular, this research only considers one attribute, therefore, we simulate X_1 only.

Notice that we consider two different choices on the amount of loan granted to borrowers. For simplicity, we first assume all borrowers will be granted with the same amount of loan L. Furthermore, we know that certain attributes determine not just whether a loan is granted but how much is lent. For example, the amount lent on a residential mortgage is determined in part by the value of the property which secures the loan. In this case, however, the amount lent will also depend on the applicants income which determines the maximum payment the borrower can make. Thus, there is a nonlinear dependence of loan amount on salary. To make a very stylized model of this non-linearity, we make our second choice to postulate a setting in which the loan granted is proportional to the square of the attribute value: $L_i = x_i \times 1,000$. Note that this model gives an incentive to lie about income not only to improve the chances of getting a loan, but also, since as the amount of lie A increases, so does the amount L_i that a bad payer can borrow, to increase the amount borrowed.

Next we discuss the effect made by the bank in verifying the correctness of attribute reported by the loan applicant. Some attributes, like the monthly salary of a salaried employee, will always be checked. However, income for restaurant staff is strongly dependent on tips which are self reported. Attributes such as this will only be checked if they seem quite unreasonable. Furthermore, we assume that only some bad payers will lie about their information as described in Section 4.2. We set X_{Bnew} to $X_B + Noi \times A(1-\eta)$.

4.4 Results and Analysis: For the Banks (Single Loan Amount)

This section analyzes the results for the banks if the amount of loan granted are the same amongst all borrowers and are fixed at $L = \$7.2K$. We will examine the revenue and profit of the banks if different values of A and η are used. In particular, we will obtain the optimal amount of effort that banks should spend on eliminating lies. We will show that under certain conditions, it is worthwhile to spend effort to eliminate lies. This depends on the strength of lies made by bad borrowers and the cost required to fix the reported value to its correct value. Using our classification model in Eq. (4.4) on past borrowers with known repayment behavior, we can divide the borrowers into four different groups:

(G_1) The classification model predicts the borrower is good and in fact she did repay her loan.

(G_2) The classification model predicts the borrower is good but in fact did not repay her loan.

(G_3) The classification model predicts the borrower is bad but she did repay her loan.

(G_4) The classification model predicts the borrower is bad and in fact she did not repay his loan.

If borrower i was classified into the group $G1$, the expected profit that the bank will earn is Q_i. If borrower i was classified into the group $G2$, the expected loss that the bank will incur is L_i. We consider n loan applicants and simulated the performance of the resulted loan portfolio using our classification model described in Eq. (4.4) for classifying good from bad borrowers. The total amount of revenue that the bank will earn as a function of η is

$$Rev(\eta) = \sum_{i=1}^{n} Q_i \mathbb{1}_{G1}(i) - \sum_{i=1}^{n} L_i \mathbb{1}_{G2}(i). \tag{4.5}$$

Here $\mathbb{1}(\cdot)$ denotes the indicator function.

Note that Eq. (4.5) indirectly captures the opportunity cost of not lending to good borrowers through the fact that the classifier balances type I (a good payer was classified as a bad payer) versus type II (a bad payer was classified as a good payer) error. If the lending criteria are too strict, $\mathbb{1}_{G1}(i)$ will usually be zero; few loans will be issued with a corresponding low sum. Notice that the borrower will lie depending on the value of η this in turn relates to total loans granted and hence revenue. When $\eta = 0$, $Rev(0)$ represents the total amount of revenue that the bank will earn when no effort has been spent to eliminate lies. Hence, the increase in revenue when the bank spent η amount of effort on eliminating lies is

$$Rev(\eta) - Rev(0). \tag{4.6}$$

Loan application cost involved in processing the application includes credit checks, property appraisals for mortgage loans or properties pledged as collateral, and basic administrative costs [2].Thus, it is clear that more effort spent on eliminating lies will result in higher costs. To simplify our model, we assumed that the total cost spent on granting loans and eliminating lies to be linear in η. We define the total cost to be $k\eta$, where k is a proportionality constant. We are interested in finding the increase in profit with respect to different levels of η, and equate it to

$$Rev(\eta) - Rev(0) - k\eta. \tag{4.7}$$

The increase in profit is the extra amount of money that banks will earn if an effort η has been spent to correct the reported attribute value X_B to a value which is closer to the attribute's true value, which we presented it as X_{Bnew}.

The effort parameter η varied from 0 to 1 with step size 0.01. Recall that $\eta = 1$ represents effort corresponding to 100% elimination of lies. We simulated 100 datasets each with 500 borrowers and reran the program 100 times. We obtained the mean and standard deviation of the average values to show the accuracy of the results. Note that, in the attempt to keep our simulation consistent with real loan portfolio, classes may have relatively few members. We did the following:

Step 1. Set η equal to 0.

Step 2. Simulated 500 borrowers using the method described in Section 4.3.

Step 3. Compute the total amount of revenue using Eq. (4.5).

Step 4. Compute the increase in revenue using Eq. (4.6).

Step 5. Compute the increase in profit using Eq. (4.7).

Step 6. Increment η by 0.01 and redo step 2 to step 5 for each dataset until η equals to 1.

Step 7. Calculate the average increase in revenue and the average increase in profit of the 100 datasets.

Step 8. Redo step 1 to step 7 100 times.

Step 9. Calculate the mean average increase in revenue and the mean average increase in profit.

μ_G	σ_G	μ_B	σ_B	n	p	I	P_N
8	1	4	1	500	0.8	20%	0.5

Table 4.1: Parameters used in Sections 4.4, 4.5 and 4.6.

4.4.1 Case I: $A = 3$, $k =$ \$390K

This case uses the parameters presented in Table 4.1, and sets $A = 3$ and $k =$ \$390K. Figure 4.1 show the mean average revenue for different levels of η, with the dashed lines showing the 95% confidence interval of the mean average values. The average revenue was calculated

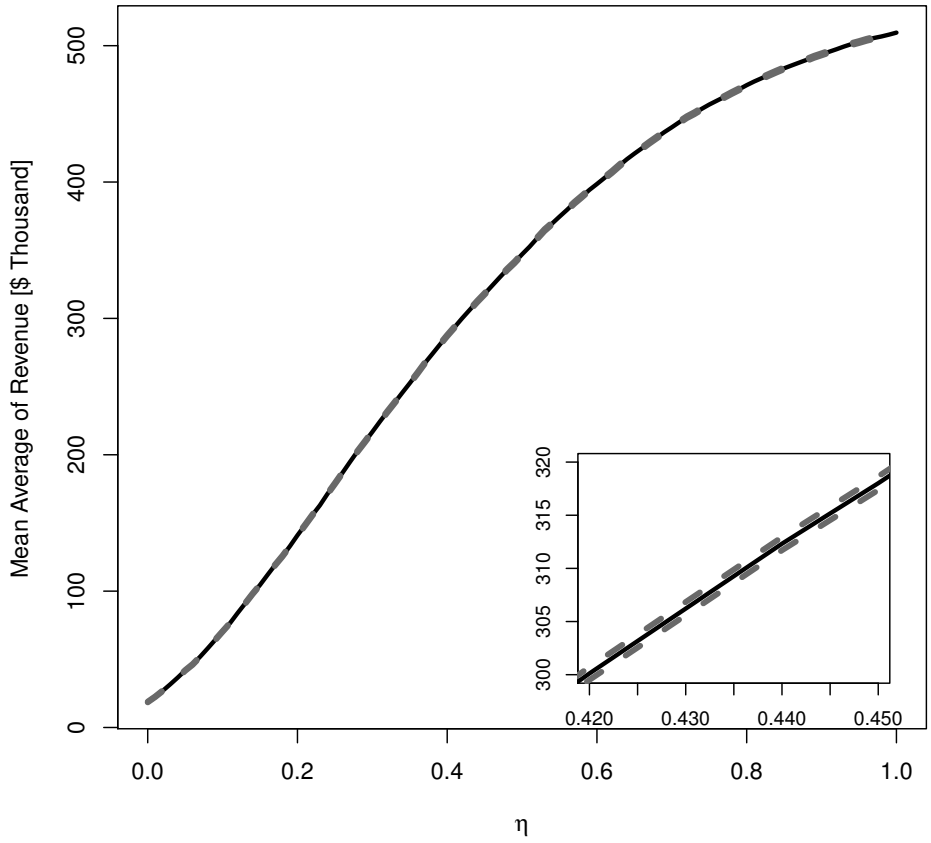


Figure 4.1: Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = 0.8$, $P_B = 0.5$ and $n = 500$ borrowers, η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 3$, $P_N = \frac{1}{2}$, $L = \$7.2K$ and $k = \$390$K to classify borrowers. We used Eq. (4.5) to calculate the revenue for each η. We computed the average revenue of 100 datasets and reran the program 100 times to obtain the average mean values of revenue, together with the 95% confidence interval of the mean values displayed in the plot as the dashed lines. The lower right corner of the plot showed a zoom-in subplot to improve the visualization of the plot.

from a random sample and, by the Central Limit Theorem [5], the sampling distribution can be approximated with a normal distribution. We can calculate the 95% confidence interval of the mean average revenue using

$$\overline{x} \pm Z^* \frac{S}{\sqrt{n^*}}, \tag{4.8}$$

where $\overline{x}$ is the mean average revenue, $Z^* = 1.96$ is the critical value for a 95% confidence interval, S is the standard deviation of the average revenues and finally n^* is the number of mean values, here equal to 100. The dashed lines in Figure 4.1 showed a very tight confidence

interval, which implies that there was a very low variability among the mean average revenues, stating that our results are very stable.

Furthermore, Figure 4.1 show a monotone increase in revenue with increasing values of η. In other words, if we put effort toward eliminating lies in the dataset, the bank will generate more revenue. In particular, if the bank devoted no effort ($\eta = 0$) to eliminate lies, the bank's revenue will equal \$18.70K, compared to if maximal effort ($\eta = 1$) was spent on eliminating all the lies from the dataset, the maximum amount of revenue that will be generated is \$509.58K. Figure 4.2 showed the mean average net increase in revenue ($Rev(\eta) - Rev(0)$), and mean average increase in profit ($Rev(\eta) - Rev(0) - k\eta$), as shown in the solid line and open square line respectively, for different levels of η. The grey dashed lines showed the 95% confidence interval of the mean values. As in Figure 4.1, the dashed lines showed a very tight confidence interval, reflecting a very low variability among the mean average values, stating that our results are very accurate. The solid line in Figure 4.2 also showed that there is a monotone increase in revenue when we increase the value of η. If $\eta = 1$, $Rev(1) - Rev(0) =$ \$490.88K, stating that if the bank devotes a full amount of effort to eliminating all the lies in the data, and if enriching the data does not bear any cost, then the bank will earn \$490.88K more. However, enriching the data requires time and effort, and it is expensive to identify liars. We have to take into account the cost of eliminating lies. This depends on both the size of lies, and the strength of effort spent on lie detection. In this case, we set $A = 3$ and $k =$ \$390K, so each unit of lies costs \$130K to eliminate. The open square line in Figure 4.2 reflected that the optimal effort (η_{opt}), that should be spent on eliminating lies is 0.67. That corresponds to generate a revenue of $Rev(\eta_{opt}) =$ \$428.85K, which requires a cost of $k\eta_{opt} =$ \$261.3K, and contributes to $Rev(\eta_{opt}) - Rev(0) - k\eta_{opt} = \$428.85 - \$18.70K - \$261.3K =$ \$148.85K increased in profit, when compared to no effort spent on eliminating lies. Notice that $\eta = 0.01$ corresponds to generate a revenue of $Rev(0.01) =$ \$22.53K, which requires a cost of $\$k \times 0.01 =$ \$3.9K, and contributes to generate a profit of $Rev(0.01) - k \times 0.01 =$ \$18.63K, which is \$68 less than when no effort was spend to identity liars. For all other values of η, it is profitable to put effort to eliminate lies.

4.4.2 Case II: $A = 3$, $k = \$1080$K

Consider the case when catching liars is very expensive. We set $k =$ \$1080K, corresponding to about \$360K to remove one unit of lies. All other parameters remain the same as in Case I. The corresponding results obtained from the mean average revenue and the mean average increase in revenue are the same as in Case I. However, since the cost to catch liars increased, the increase in profit will be different. Figure 4.3 shows the plot of mean average increase in

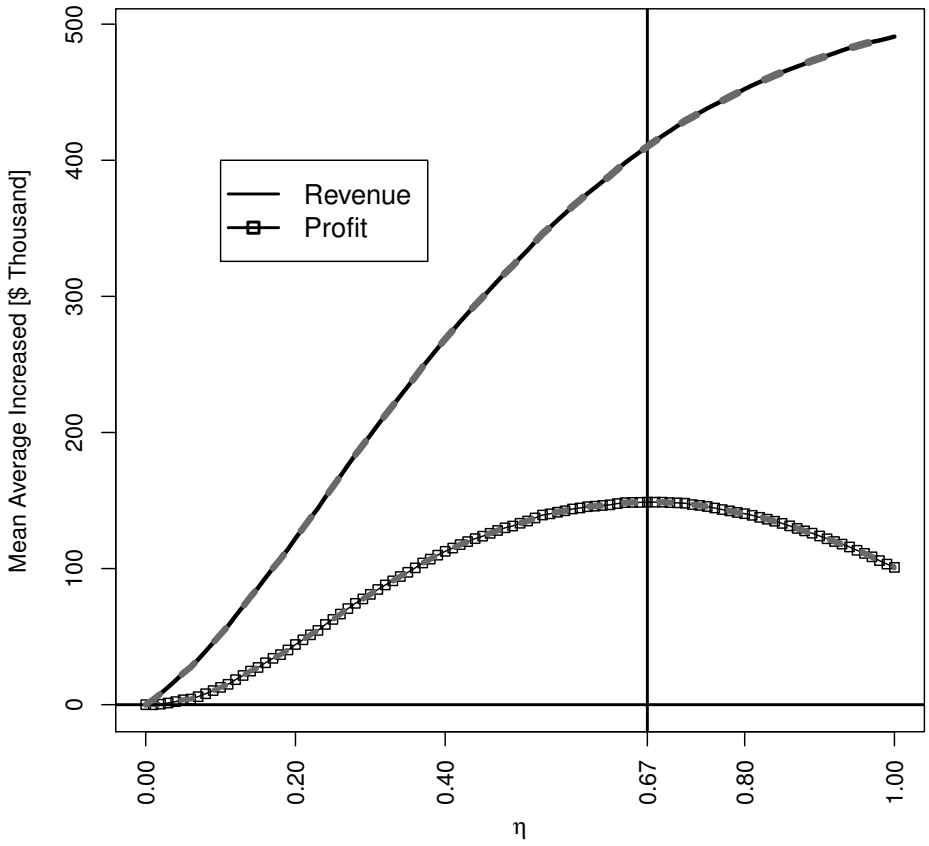


Figure 4.2: Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = 0.8, P_B = 0.2$ and $n = 500$ borrowers,η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 3$, $L = \$7.2\text{K}$, $k = \$390\text{K}$ and $P_N = \frac{1}{2}$ to classify borrowers. We used Eqs. (4.5), (4.6) and (4.7) to calculate the increase in revenue and profit for each η. We reran the program 100 times to obtain the average mean values of increase in revenue and profit as stated in the solid and open square lines respectively. The grey dashed lines displayed the 95% confidence interval of the mean values.

revenue (solid line) and profit (open square line) versus different levels of η, with the dashed lines showing the 95% confidence interval of the mean values. This figure provided evidence that for all values of η, the corresponding increase in net profit will be negative, stating that it is too expensive to eliminate lies. Thus, banks should not spend effort to catch these liars.

4.4.3 Case III: $A = 4$, $k = \$520\text{K}$

If borrowers can choose how much to lie, and if they assume that the bank is not checking their lies ($\eta = 0$), they might lie so as to completely mimic good borrowers. Using the parameters of

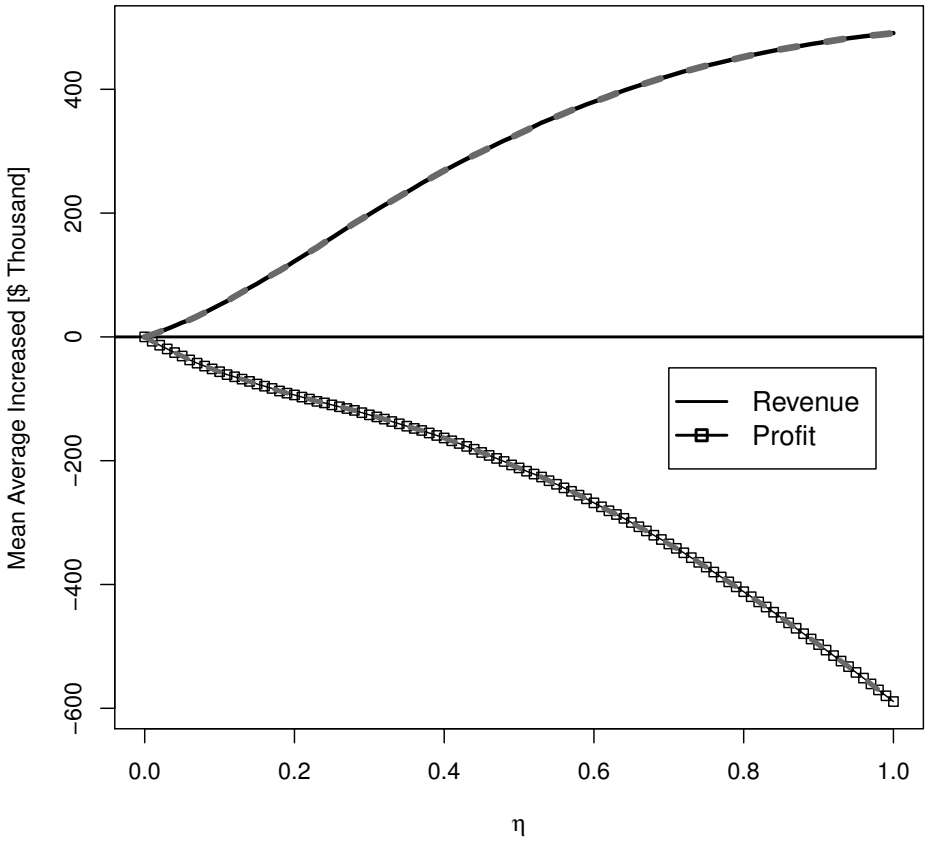


Figure 4.3: Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = 0.8, P_B = 0.5$ and $n = 500$ borrowers, η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 3$, $L = \$7.2\text{K}$, $k = \$1080\text{K}$ and $P_N = \frac{1}{2}$ to classify borrowers. We used Eqs. (4.5), (4.6) and (4.7) to calculate the increase in revenue and profit for each η. We reran the program 100 times to obtain the average mean values of increase in revenue (solid line) and profit (open square line), together with the 95% confidence interval of the mean values, displayed as the dashed lines.

Table 4.1 in which $\mu_B = 4$ and $\mu_G = 8$, this suggests liars might choose $A = 4$.

In this case it would be impossible to distinguish good borrowers from bad borrowers who lie. But it remains possible to distinguish the group containing both good borrowers and bad borrowers who lie from the group of truthful borrowers with bad credit prospects. Even assuming that the bank could perfectly distinguish between these groups, it would still profit from just 2/3 of its loans. (This is, of course, an upper bound to actual performance.) This is because it would lend to all the good borrowers (half the applicants) and all the bad liars (half of the remaining 50% of applicants). So the bank would lend to 75% of the borrowers, only 50% of

which would be good borrowers. Sadly, the bank only charges an interest rate of 20%, so it loses 5 times as much on each bad liar as it makes on each good borrower, causing it to lose money on this loan book. At this point it would decide not to lend money at all.

It remains suboptimal to lend money at all, even with no cost of checking, until $\eta = 0.25$ (as seen by the fact that Figure 4.4 has no increase in revenue as a function of η until $\eta = 0.25$). At that stage the cost of checking the applicants is larger than the revenue earned and the net profit is negative. We have to wait until $\eta = 0.55$ where the cost of checking equals the profit made; further improvement in checking continues to yield benefits until $\eta = 0.76$ when the classifier is nearly perfect at distinguishing liars from truthful good borrowers. After this point diminishing returns in the profitability in lending mean that the added cost in lending is not worth it.

4.4.4 Case IV: $A \in \{1, 1.25, 1.5, ..., 6\}$, Unit Cost = \$130K

In order to obtain deeper insights about the model's performance, we varied A from 1 to 5 with step size 0.25. To keep results comparable with case I and III, we assume that each lie unit costs \$130K to remove. For each value of A, we computed the corresponding values of k, followed step 1 to step 5 discussed at the beginning of this section, and obtained η_{opt}. Tables 4.2 displayed different values of A with the corresponding values of k, η_{opt}, $A(1-\eta_{opt})$, required cost and maximum profit. Notice that required cost = $k\eta_{opt}$, and

$$\text{maximum profit} = Rev(\eta_{opt}) - k\eta_{opt}.$$

Recall that A is the level of lies that bad borrowers added to their attribute values, and $A(1-\eta)$ is the actual amount of lies remaining in X_{Bnew} after accounting for the cost of reducing the lies. Table 4.2 shows that for values of A between 1 to 4.25, the model suggested to set η_{opt} to a value which reduced the difference between the reported and true value of the attribute by 1. This explains the reason for having all the values in the column "$A(1-\eta_{opt})$" in Tables 4.2 to be approximately equal to 1. As A increases, the required cost to check for lies increase and leads to lower profit. When $A = 3.75, 4$ or 4.25 and if lenders did not spend any effort to eliminate lies, the resulted revenue will be zero, and our model suggested lenders to set their effort to η_{opt} so as to generate some profit. For $A = 4.5, 4.75, 5$ or 5.25, there will only be negative profit generated for all values of η, stating that it is not profitable to grant money. To illustrate this phenomenon, we set A to 4.5 to generate Figure 4.5, which shows that the entire profit curve is below zero. The corresponding cost required to identify liars is too high, and our model suggest not to do business. Particularly, if we set $A = 4.5$ and alter the unit cost required to check for lies, we realized that if each unit of lies costs higher than \$122K to eliminate, then it is not profitable to do business. When $A = 5.5, 5.75$ or 6, the lies are more than necessary,

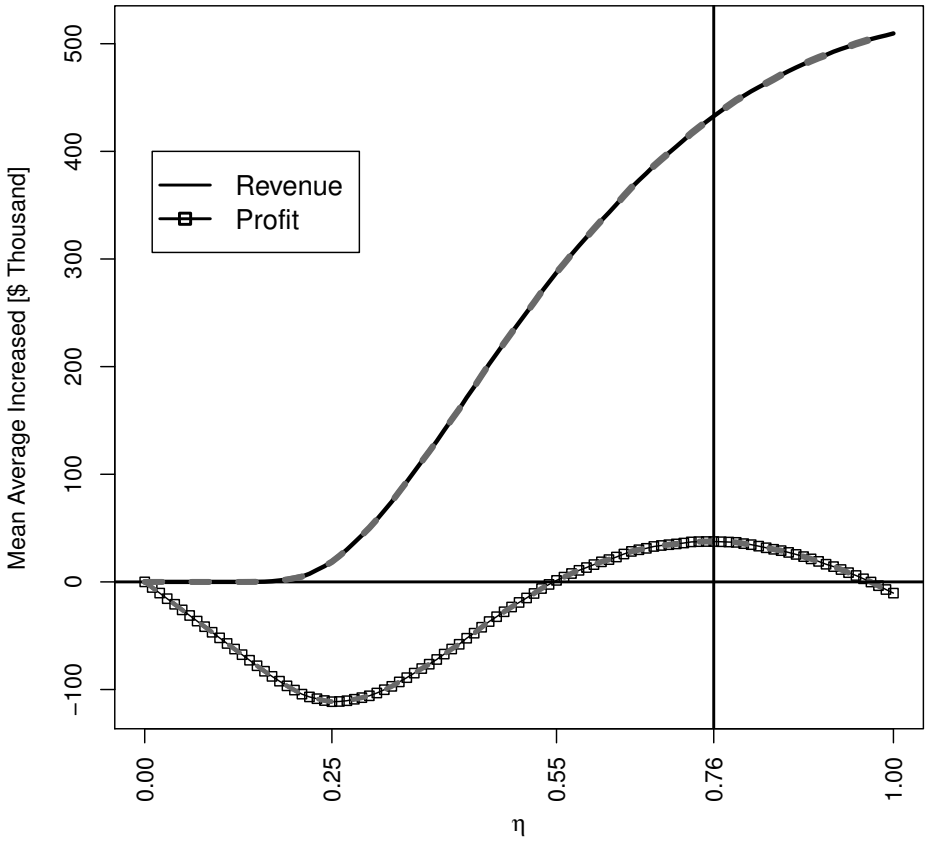


Figure 4.4: Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = 0.8, P_B = 0.2$ and $n = 500$ borrowers, η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 4$, $k = \$520\text{K}$, $L = \$7.2\text{K}$ and $P_N = \frac{1}{2}$ to classify borrowers. We used Eqs. (4.5), (4.6) and (4.7) to calculate the increase in revenue and profit for each η. We reran the program 100 times to obtain the average mean values of increase in revenue and profit, together with the 95% confidence interval of the mean values, displayed as the dashed lines.

paradoxically making them easier to detect again. The classification model itself works well enough to distinguish bad liars and does not require additional effort spend on checking lies. Figure 4.6 depicts the optimal effort η_{opt} which the banks should exert to correspond to lies of level A by the bad liar loan applicants. This curve is increasing in A if bad liars increase their lies, the banks will also need to devote more efforts to checking for lies. However, the effort seems to flatten out with increased lie level (the second derivative of the curve is negative), perhaps because diminishing returns are met once most liars make it through the screens.

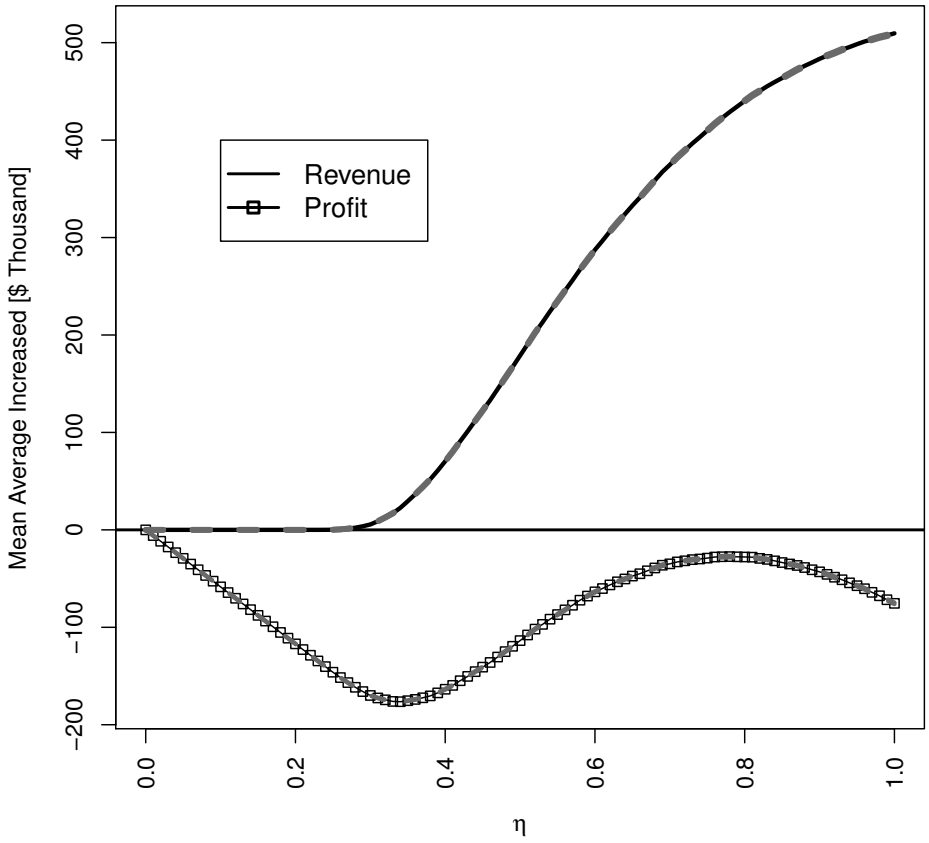


Figure 4.5: Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = 0.8, P_B = 0.2$ and $n = 500$ borrowers, η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 4.5$, $k = \$520$K, $L = \$7.2$K and $P_N = \frac{1}{2}$ to classify borrowers. We used Eqs. (4.5), (4.6) and (4.7) to calculate the increase in revenue and profit for each η. We reran the program 100 times to obtain the average mean values of increase in revenue and profit, together with the 95% confidence interval of the mean values, displayed as the dashed lines.

4.4.5 Case V: $A \in \{1, 1.25, 1.5, ..., 5\}$, Unit Cost = \$260K

We double the unit cost required to check for borrowers' lies and assume that each unit costs \$260K to remove. Table 4.3 shows that for all values of A, it is not optimal for lenders to put effort to check for lies. When A is smaller than 3.5, lenders will still be profitable to lend money, since the classification model in Eq. (4.4) provides very accurate results. For values of A greater that 3.5, our model suggest not to do business. Note that lenders can decide the amount of effort spend to check for lies, that was indeed affected by its cost, while bad liars can determine the amount of lies to add onto the true attributes. A more detailed investigation

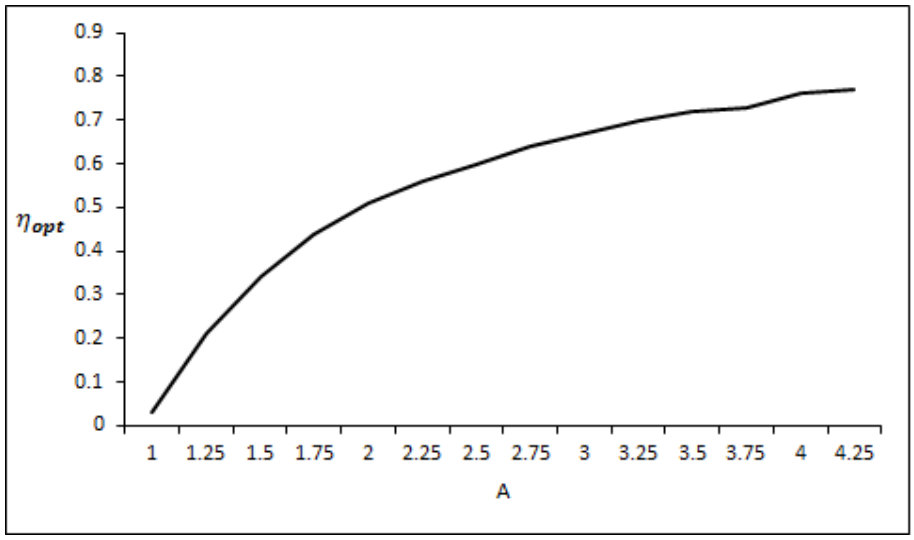


Figure 4.6: The plot of η_{opt} versus A from Table 4.2.

A	k	η_{opt}	$A(1-\eta_{opt})$	Required Cost	Rev(0)	Maximum Profit
1	\$130,000	0.03	0.97	\$3,900	\$427,496	\$427,603
1.25	\$162,500	0.21	0.99	\$34,125	\$391,135	\$395,064
1.5	\$195,000	0.34	0.99	\$66,300	\$347,001	\$362,554
1.75	\$227,500	0.44	0.98	\$100,100	\$297,778	\$330,087
2	\$260,000	0.51	0.98	\$132,600	\$240,915	\$297,587
2.25	\$292,500	0.56	0.99	\$163,800	\$178,849	\$265,054
2.5	\$325,000	0.6	1.00	\$195,000	\$116,724	\$232,496
2.75	\$357,500	0.64	0.99	\$228,800	\$59,808	\$200,054
3	\$390,000	0.67	0.99	\$261,300	\$18,703	\$167,554
3.25	\$422,500	0.7	0.98	\$295,750	\$1,906	\$134,987
3.5	\$455,000	0.72	0.98	\$327,600	\$2	\$102,587
3.75	\$487,500	0.73	1.01	\$355,875	\$0	\$70,030
4	\$520,000	0.76	0.96	\$395,200	\$0	\$37,505
4.25	\$552,500	0.77	0.98	\$425,425	\$0	\$5,060
4.5	\$585,000	0	4.50	\$0	\$0	\$0
4.75	\$617,500	0	4.75	\$0	\$0	\$0
5	\$650,000	0	5.00	\$0	\$0	\$0
5.25	\$682500	0	5.25	\$0	\$0	\$0
5.5	\$715000	0	5.50	\$0	\$46,410	\$46,410
5.75	\$747500	0	5.75	\$0	\$129,437	\$129,437
6	\$780000	0	6.00	\$0	\$194,511	\$194,511

Table 4.2: Comparison of results generated using different values of A for L = \$7.2K and each unit of lies costs \$130K to remove.

of the game between the borrowers and lenders, which takes into account of the unit cost of checking lies and the unit cost to make a lie, is the main topic for next chapter.

A	k	η_{opt}	$A(1-\eta_{opt})$	Required Cost	Rev(0)	Maximum Profit
1	\$260,000	0	1.00	\$0	\$427,496	\$427,496
1.25	\$325,000	0	1.25	\$0	\$391,135	\$391,135
1.5	\$390,000	0	1.50	\$0	\$347,001	\$347,001
1.75	\$455,000	0	1.75	\$0	\$297,778	\$297,778
2	\$520,000	0	2.00	\$0	\$240,915	\$240,915
2.25	\$585,000	0	2.18	\$0	\$178,849	\$178,849
2.5	\$650,000	0	2.50	\$0	\$116,724	\$116,724
2.75	\$715,000	0	2.75	\$0	\$59,808	\$59,808
3	\$780,000	0	3.00	\$0	\$18,703	\$18,703
3.25	\$845,000	0	3.25	\$0	\$1,906	\$1,906
3.5	\$910,000	0	3.50	\$0	\$2	\$2
3.75	\$975,000	0	3.75	\$0	\$0	\$0
4	\$1,040,000	0	4.00	\$0	\$0	\$0
4.25	\$1,105,000	0	4.25	\$0	\$0	\$0
4.5	\$1,170,000	0	4.50	\$0	\$0	\$0
4.75	\$1,235,000	0	4.75	\$0	\$0	\$0
5	\$1,300,000	0	5.00	\$0	\$0	\$0

Table 4.3: Comparison of results generated using different values of A for L = \$7.2K and each unit of lies costs \$260K to remove.

4.5 Results and Analysis: For the Banks (Attribute Dependent Loan Amount)

It is interesting to consider the case where the amount of loan granted is somehow dependent on the value of the attribute of the corresponding borrower. To investigate this, we consider $L_i = \$x_i \times 1{,}000$. Notice that $X_G \sim N(8,1), X_B \sim N(4,1), P_G = 0.8$ and $P_B = 0.2$, therefore, $E(X) = 8 \times 0.8 + 4 \times 0.2 = 7.2$ and so this explains the reason for choosing a loan amount of \$7.2K in Section 4.4. If there are no liars in the lending business, the potential amount of loan granted to all borrowers in both single loan amount and attribute dependent loan amount will roughly be the same, at around \$7.2K $\times$500 = \$3.6MM. In this section, we keep all the conditions same as in Section 4.4, and only change the amount granted to each borrower to $L_i = x_i \times 1{,}000$. We compare and analyze the difference between granting at a single loan amount and at an attribute dependent loan amount in some of the cases described as in Section

4.4. We will not repeat the discussion in cases which generates similar results as in Section 4.4.

4.5.1 Case I: A = 3, k = $390K

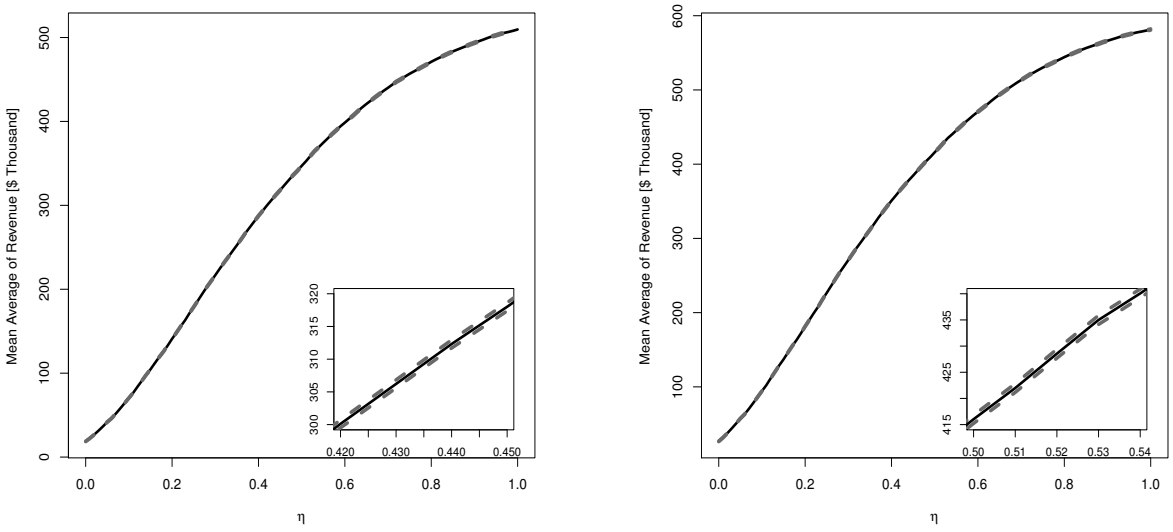
In section 4.4, we fixed the amount of loan granted to be the same amongst all borrowers at $7.2K. Using the parameters as stated in Table 4.1 and setting $A = 3$, the potential amount of loan grant to good payers as $7.2K $\times 500 \times 0.8$ = $2.88MM, to bad true tellers and bad liars are the same as $7.2K $\times 500 \times 0.2 \times 0.5$ = $360K. Therefore, the total potential amount of loans granted is $3.6MM, which compose of 80% to good payers, 10% to bad true tellers and another 10% to bad liars. By setting $L_i = \$x_i \times 1{,}000$, we can roughly estimate the potential amount of loans granted to all good payers as $8K $\times 500 \times 0.8$ = $3.2MM, to bad honest payers as $4K $\times 500 \times 0.2 \times 0.5$ = $200K and to bad liars as $7K $\times 500 \times 0.2 \times 0.5$ = $350K. The total potential amount of loans granted is $3.75MM, of which 85% to good payers, $5.\dot{3}$% to bad true tellers and $9.\dot{3}$% to bad liars. Varying the loan amount granted by borrowers' attribute value increases the proportion of money granted to good payers and decreases the proportion of money granted to bad payers. Comparing the two choices of L and taking into account of the three units of lies added by bad liars, the attribute dependent scenario increased the potential amount of revenue from good payers by $\frac{\$3.2-\$2.88}{\$2.88} \times 100\% = 11.11\%$, while decreased the potential amount of loss to bad true tellers by $\frac{\$360-\$200}{\$360} \times 100\% = 44.44\%$, and to bad liars by $\frac{\$360-\$350}{\$360} \times 100\% = 2.77\%$.

Table 4.4 compared the banks' revenue and profit of the two different choices of L. It shows that setting the amount of loan granted to be attribute dependent will generate more revenue and hence more profitable. This result is consistent with the fact that more loans will be allocated to good payers in the attribute dependent scenario. Although, our model suggests a higher η_{opt} for the attribute dependent scenario, which indeed increases the cost to check for lies by $0.02 \times \$390K = \$7.8K$, the mean average increase in profit is 43.82% higher than the single loan amount scenario. Figure 4.7 showed a side by side plot of the mean average revenue($Rev(\eta)$) for different levels of η and for the two different choices of L, with the dashed lines showing the 95% confidence interval of the mean average values. Both of the plots exhibit very similar trend, with the attribute dependent scenario generates higher revenue. Figure 4.8 showed the mean average net increase in revenue ($Rev(\eta) - Rev(0)$), and the mean average increase in profit ($Rev(\eta) - Rev(0) - k\eta$) as shown in the solid line and the open square line respectively, for different levels of η and for the two different choices on loan amount granted. Both of the curves show that the mean average increase in revenue and profit shifted up in the attribute dependent loan scenario. In particular, when η equals to 0.01, the constant loan scenario generates negative profit, due to the fact that the cost required to check for lies is higher

than revenue, while in the attribute dependent loan scenario, positive profit was generated for all values of η.

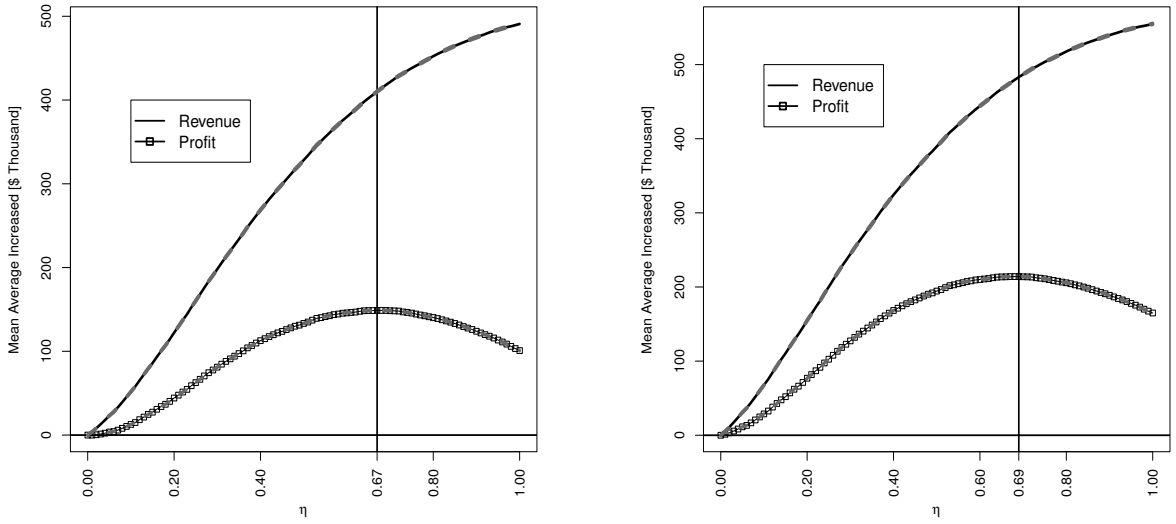
	$L = \$7.2\text{K}$	$L_i = \$x_i \times 1{,}000$	Percentage Increased
$Rev(0)$	\$18.07K	\$26.53K	46.82%
$Rev(1)$	\$509.58K	\$581.46K	14.11%
$Rev(1) - Rev(0)$	\$490.88K	\$554.93K	13.05%
η_{opt}	0.67	0.69	2.99%
$Rev(\eta_{opt})$	\$428.85K	\$509.7K	18.85%
$k\eta_{opt}$	\$261.3K	\$269.1K	2.99%
$Rev(\eta_{opt}) - Rev(0) - k\eta_{opt}$	\$148.85K	\$214.07K	43.82%

Table 4.4: Comparison of the bank's revenue and profit on the two different choices of loan amount: (a) $L = \$7.2\text{K}$ and (b) $L = \$X_i \times 1{,}000$.



(a) $L = \$7.2\text{K}$ (b) $L_i = \$x_i \times 1{,}000$

Figure 4.7: Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = 0.8, P_B = 0.2$ and $n = 500$ borrowers, η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 3$, $P_N = \frac{1}{2}$ and $k = \$390\text{K}$ to classify borrowers. We used Eq. (4.5) to calculate the revenue for each η. We computed the average revenue of 100 datasets and reran the program 100 times to obtain the average mean values of revenue, together with the 95% confidence interval of the mean values displayed in the plot as the dashed lines. The lower right corner of the plot showed a zoom-in subplot to improve the visualization of the plot.



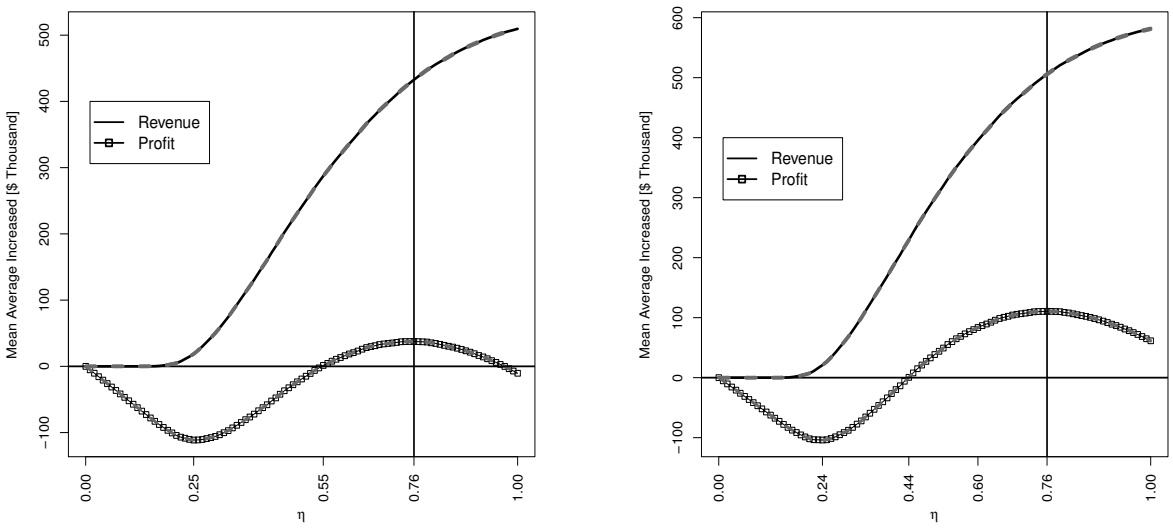
(a) $L = \$7.2\text{K}$ (b) $L_i = \$x_i \times 1{,}000$

Figure 4.8: Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = 0.8, P_B = 0.2$ and $n = 500$ borrowers, η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 3$, $k = \$390\text{K}$ and $P_N = \frac{1}{2}$ to classify borrowers. We used Eqs. (4.5), (4.6) and (4.7) to calculate the increase in revenue and profit for each η. We reran the program 100 times to obtain the average mean values of increase in revenue and profit as stated in the solid and open square lines respectively. The grey dashed lines displayed the 95% confidence interval of the mean values.

4.5.2 Case III: A = 4, k = $520K

In this case, bad liars set A to 4 and make them indistinguishable from good payers, the potential amount of loan granted to good payers and bad true tellers will be the same as in case I in this section. However, the potential amount of loan granted to bad liars will increase from \$350K to \$400K. Since in this case, bank would lent 94.74% of loans to potential good payers, while only 84.21% of them are actually good payers and 10.53% of them are bad liars. For the two different choices of L, Figure 4.9 shows the mean average increased in revenue and profit. In both scenarios, our model suggest to set η_{opt} to 0.76, however, in the attribute dependent loan scenario, the increase in profit $(Rev(\eta_{opt}) - Rev(0) - k\eta_{opt})$ is \$110.55K, which is 194.72% more than the single amount loan scenario at \$37.51K.

Notice that case II, IV and V generate similar results using attribute dependent loan amount and single loan amount. We will not repeat the analysis and discussion here.



(a) $L = \$64$ MM

(b) $L_i = \$x_i^2 \times 1{,}000$

Figure 4.9: Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = P_B = 0.5$ and $n = 500$ borrowers, η varied from 0 to 1 with step size 0.01. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $A = 4$, $k = \$520$MM and $P_N = \frac{1}{2}$ to classify borrowers. We used Eqs. (4.5), (4.6) and (4.7) to calculate the increase in revenue and profit for each η. We reran the program 100 times to obtain the average mean values of increase in revenue and profit, together with the 95% confidence interval of the mean values, displayed as the dashed lines.

4.6 For the borrowers:

Here we show that, provided liars know the attributes being checked and provided a linear lie elimination cost model as above, liars can always adjust the amount of lies A so as to escape from credit checks. We varied A from 0 to 12 with step size 0.04 to capture several different system behaviors. For each value of A, we fixed η and used the classification model in Eq. (4.4), with parameters as stated in Table 4.1 to generate Figure 4.10 and Figure 4.11. Recall from Section 4.4 that G_2 represents the group of borrowers who received a loan but did not repay, while G_4 represents the group of borrowers who correctly did not obtain a loan because they would not repay. Using the information in Table 4.5, we can calculate odds ratio [15] to determine the effect of bad borrower lying towards getting loans. From Lohr [17], we know that odds ratio is the ratio of $\frac{a}{b}$ to $\frac{c}{d}$, which is $\frac{a/b}{c/d}$, $\frac{a}{b}$ in turn gives the ratio of getting credit to not getting credit for bad borrowers who lie, $\frac{c}{d}$ gives the same ratio for bad borrowers who tell the truth. If this ratio is bigger than one, bad borrowers should lie to guarantee the best odds of getting a loan. Of course bad borrowers can only implement this if they know the

values of the parameters in Table 4.1 as well as those selected by the lender; whether such knowledge is realistic is open for debate. That the prospective borrower have this information represents a worst case for the lender. The strategy of borrowers depend on odds ratio, in addition, our classification model as presented in Eq. (4.4) is independent of the loan amount granted. Therefore, no matter L_i is fixed at \$7.2K or dependents on the borrowers attribute $\$x_i \times 1,000$, the results for the borrowers will be the same. We will not repeat the analysis for borrowers in this section. However, in the next chapter, we will introduce a measure on the utility of borrowers which is correlated to the loan amount granted.

We simulated 500 borrowers and computed the average odds ratio of 100 datasets. We set $\eta = 0$, 0.2 and 0.4 to compare the effect of different amount of effort spent on eliminating lies with the level of odds ratio that can be obtained. Figure 4.10 shows the plot of average odds ratio using different values of η for different values of A. Note that we only considered non-trival odds ratios, ignoring the data when either "b" or "c" or both "b" and "c" in Table 4.5 is zero and this contributes to the discontinuity observed in the curves shown in Figure 4.10. Furthermore, Figure 4.10 shows that higher values of η correspond to higher values of A in order to obtain the same level of odds ratio. That is to say, if banks put more effort on eliminating lies, bad borrowers must increase the size of their lies in order to maintain the same chances to be classified into the group of good borrowers. Note that this result is seems strange if we consider the massive lie required to get past $\eta = 0.99$ level checking; remember that our model does not consider the original attribute reported and disqualify applicants for lying, we only disqualify applicants for low corrected scores. Moreover, bad borrowers who decided to lie in order to obtain credit should use the values A where odds ratio is greater than 1.

In order to obtain the confidence interval estimate of average odds ratio, we set $\eta = 0.2$ and varied A from 0 to 12. We simulated 100 datasets each with 500 borrowers and reran the program 100 times. For each A, we computed and sorted the 100 average odds ratio. Then, we found the 5^{th} and 95^{th} average odds ratio, which gave the 95% confidence interval of mean odds ratio. Figure 4.11 shows the plot of the mean average of odds ratio and the confidence interval of different values of A. It also shows that extremely large lies ($A > 9.84$) are counter productive, and that very small lies ($A < 0.28$) are not worth the effort. This shows that the behavior of a borrower that wishes to lie in a credit check has to be strategic, and therefore it will take some effort from the bank's point of view to detect them.

4.7 Conclusions

Improper handling of credit fraud can cause inaccurate loan decisions which contribute to suboptimal loan book profitability. Bad borrowers may exaggerate (i.e. lie about) one or more

	G_2	G_4
Lied	a	b
Didn't Lie	c	d

Table 4.5: A two by two table used to calculate odds ratio to determine whether bad borrowers should lie in order to increase their chance of getting loans.

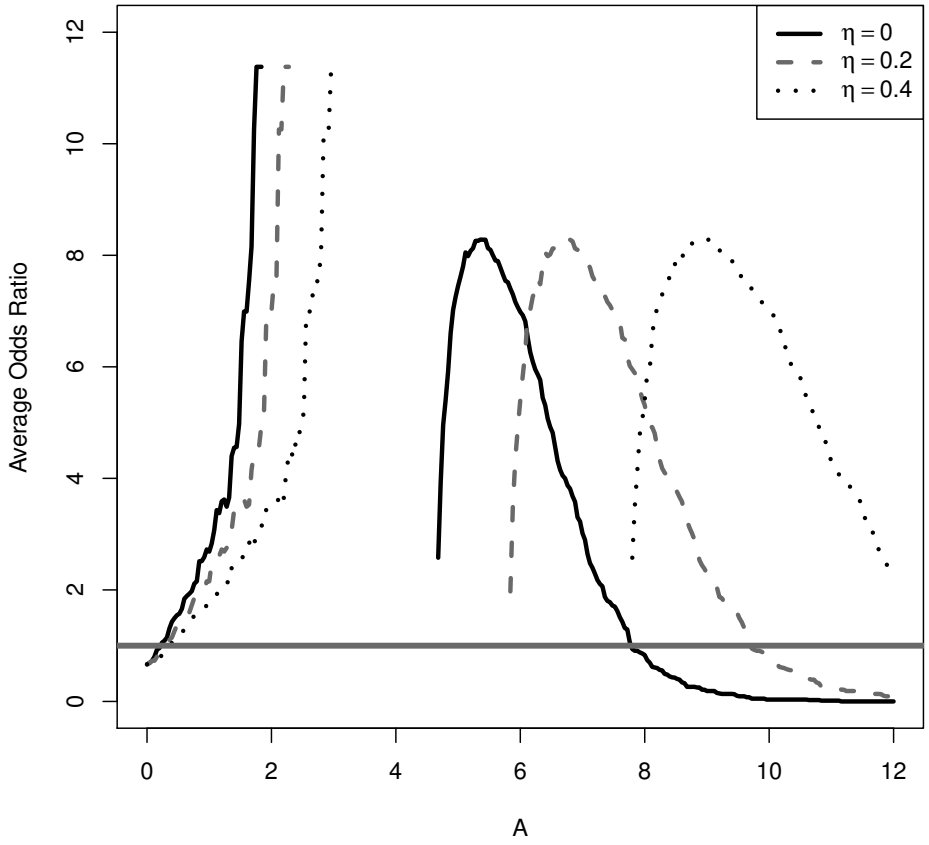


Figure 4.10: Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = P_B = 0.5$ and $n = 500$ borrowers. We varied A from 0 to 12 with step size 0.04. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$ and $P_N = \frac{1}{2}$ to generate the table shown in Table 4.5. To compare the effect of η, we set η equals to 0, 0.2 and 0.4. For each value of η, we computed the average odds ratio of 100 datasets. The above plot showed that bad borrowers should lie in the values of A where odds ratio is greater than 1, in order to increase their chance of having loan approval. Furthermore, it showed that increasing the value of η will require a higher value of A to maintain the same level of odds ratio.

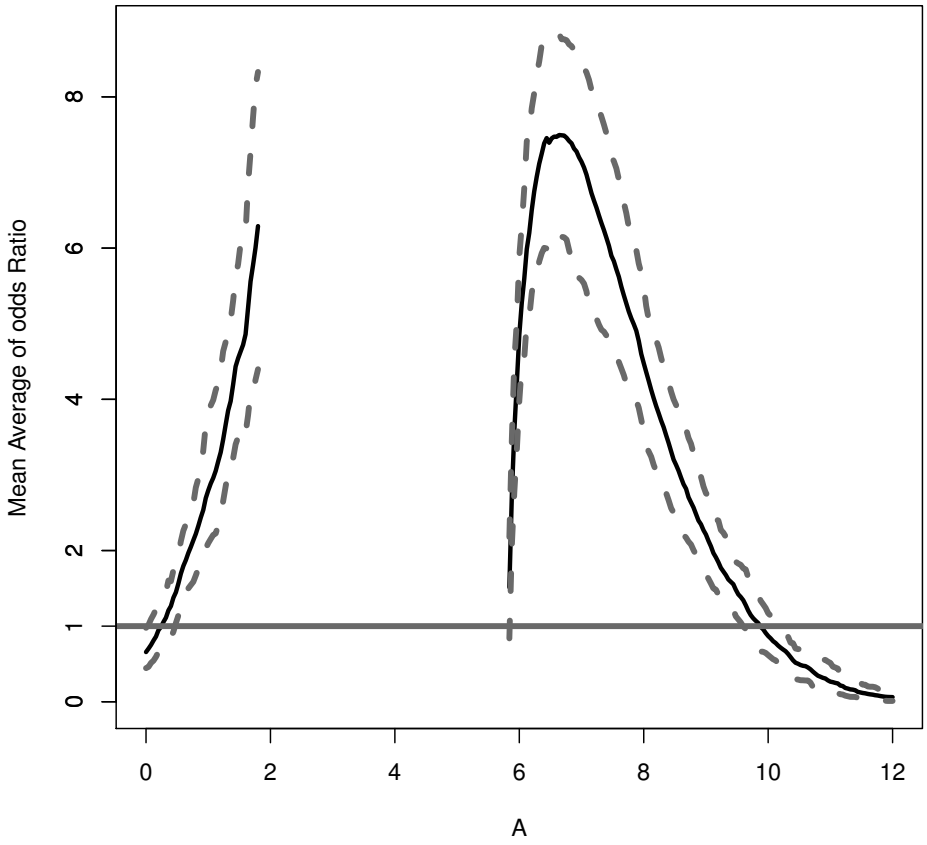


Figure 4.11: Data was simulated from $X_G \sim N(8,1)$ and $X_B \sim N(4,1)$ with $P_G = P_B = 0.5$ and n = 500 borrowers. We varied A from 0 to 12 with step size 0.04. We applied the set condition of Eq. (4.4) with initial values, $I = 20\%$, $\eta = 0.2$ and $P_N = \frac{1}{2}$ to generate the table shown in Table 3. We reran the program for 100 times to obtain the mean average of odds ratios. We found the confidence interval estimates of the average odds ratio, as shown as the dashed lines in the above plot.

of their attributes in order to obtain credit. In this chapter we presented a simple stylized model of this phenomenon to investigate how banks might respond to these lies. In particular, we studied the difference between granting a single loan amount versus an attribute dependent loan amount. The overall results of these two different choices are roughly the same, which gives us support that our analysis are "robust" to model specification.

From the bank's perspective our study shows that while there are cases in which huge lies are better left alone, intelligent effort spent on reducing lies usually increases revenue. Whether this increased revenue corresponds to increased profit depends, of course, on the cost spent to reduce the lies and the degree to which bad borrowers lie. In a real-world setting, the cost to

banks comprises the cost of carrying out credit checks to verify the correctness of the attribute values reported in the application form, and purchasing credit history and FICO scores of loan applicants from the credit bureau. To cover the cost, banks can impose an up-front fee to loan applicants to have their loan established. However, to secure business, many banks will waive this application fee for their customers, instead incorporating it to the cost of loans through the interest rate charged. The example discussed in this chapter suggests that, if eliminating lies is not too expensive, a good strategy for a bank is to reduce the impact of lies only until they are small enough that the classifier can take over and itself distinguish good from bad borrowers. In this setting, very small lies are not worth removing while huge, ridiculous, lies may actually help the classifier, always assuming that only bad borrowers lie. Of course, the threshold between 'small', 'normal' and 'huge' need to be determined using the appropriate classifier technique.

From the borrower's point of view, we showed that it is possible for bad borrowers to lie intelligently to pass credit checks and obtain loans. In particular, if the bad borrower is aware of the attribute and the type of model used to classify prospective borrowers as well as the bank's cost of reducing lies, we show that it is possible for this clever bad borrower to select a lie of optimal size to increase their chance of receiving credit while being small enough that the bank's credit easily deleted them. Moreover, in the case the loan amount granted depends on borrower's attribute, the amount of the lie affects the size of the loan that a bad borrower would be granted. The optimum lie is therefore the result of a tradeoff between the chances of getting a loan and the size of the loan so obtained.

In conclusion, the detection of lies in statistical classifier input data can bring important gains for the bank. The problem is very complex, since liars have a strong incentive to behave strategically. More research should be focused on identifying incorrect attributes. It would be interesting to study how robust these results are in a model in which good borrowers as well as bad ones lie, perhaps for reasons external to the loan-granting process. Another alternative might be to attempt not to correct the attributes stated on forms as modeled here but to identify liars and simply reject their applications.

Acknowledgements

C. Bravo acknowledges the support of CONICYT FONDECYT Initiation into Research Grant Nr. 11140264 and the Institute of Complex Engineering Systems (ICM: P-05-004-F, CONICYT: FBO16).

M. Davison acknowledges the support of the NSERC Discovery Grant and the Canada Re-

search Chairs programs.

Bibliography

[1] M. Chong, C. Bravo and M. Davison, How much effort should be spent to detect fraudulent applications when engaged in classifier-based lending?, *Intelligent Data Analysis*, **19**(s1), (2015), S87-S101

[2] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens and J. Vanthienen, Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, **54**(6), (2003), 627-635.

[3] B. Liu, E. Roca, What drives mortgage fees in Australia?, *Accounting & Finance*, (2014). Accepted for publication. Available Online: DOI: 10.1111/acfi.12068.

[4] C. Bravo, L. Thomas and R. Weber, Improving credit scoring by differentiating defaulter behavior, *Journal of the Operational Research Society*, **66**(5) (2014), 771-781.

[5] De Veaus, Velleman, Bock, Vukov and Wong, *Stats Data and Models*, Pearson Canada, 2012.

[6] E.I. Altman, Financial Ratios, Discriminant analysis and the prediction of corporate bankruptcy, *The Journal of Finance*, **23**(4) (1968), 589-609.

[7] J. Graham, S. Li and J. Qiu, Corporate misreporting and bank loan contracting, *Journal of Financial Economics* **89**(1), (2008), 44-61.

[8] J. Griffin and G. Maturana, Who Facilitated Misreporting in Securitized Loans?, *The Journal of Finance*, (2015). Accepted for Publication: Available online. DOI: 10.1111/jofi.12255.

[9] K. Dejaeger, B. Hamers, J. Poelmans and B. Baesens, A novel approach to the evaluation and improvement of data quality in the financial sector, *Proceedings of the International Conference on Information Quality (ICIQ 2010)*, Little Rock, AR, United States, 2010.

[10] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, Lyn C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research* **247**(1), (2016), 124-136.

[11] L. Guiso, P. Sapienza and L. Zingales, The determinants of attitudes towards strategic default on mortgages, *The Journal of Finance*, **68**(4), (2013), 1473-1515.

[12] L. Thomas, A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers, *International Journal of Forecasting*, **16**(2), (2002), 149-172.

[13] L. Thomas, D. B. Edelman and J. N. Crook, *Credit Scoring and Its Applications*, SIAM, Philadelphia, USA, 2002.

[14] M. Garmaise, Borrower Misreporting and Loan Performance, *The Journal of Finance*, **70**(1), (2015), 449-484.

[15] M.M. Triola and M.F. Triola, Biostatistics for the biological and health sciences, Pearson Addison-Wesley, Boston, USA, 2006, pp. 130-134.

[16] R. Anderson, *The Credit Scoring Toolkit: Theory and practice for retail credit risk management and decision automation*, Oxford University Press, Oxford, UK, 2007.

[17] S.L. Lohr, Sampling Design and Analysis, Brooks/Cole, Boston, USA, 1999, pp. 320-321.

Chapter 5

Cost Effective Game of Banks and Dishonest Borrowers

5.1 Introduction

In recent years the application of machine learning or classifiers on retail and small commercial lending has become more and more popular. Many classifier algorithms are used for credit scoring, as reviewed in Baesens *et al*. [1] and in Lessmann *et al*. [5]. According to Thomas *et al*. [7] credit scoring refers to the use of a numerical tool to rank borrowers according to their desirability as loan counterparty. This desirability reflects their probability of making repayments as well as the fraction of a loan expected to be recovered in the event of a default. To build a scoring model, developers analyze historical data on the performance of previously made loans to determine which borrower characteristics are useful in predicting whether the loan performed well. From this analysis the next step is to build a credit scoring model, which is a mathematical formula that can be used to combine the value of the selected characteristics to generate a credit score. A common credit score known as the FICO [1] score, is a measure of consumer credit risk. This score ranges from 300 to 850; the higher the score the more creditworthy the applicant is deemed to be. As mentioned in Frankel [9], lower credit scores are systematically associated with a higher probability of default on mortgage loans.

To apply for a loan, prospective borrowers have to provide their attributes by filling out an application form. The credit analyst will then input the required attributes into the credit scoring model to generate a credit score. To balance the risk and return of the financial institution, a pre-determined cutoff score will be used to obtain the granting decision. Applicants with scores higher than the cutoff score will be granted credit; while applicants with scores lower than the

[1]The acronym FICO is derived from "Fair Issac & Co". This firm is a leading analytics software company which produces statistical models to predict consumer behavior.

cutoff score will be denied credit. A Credit scoring model can correctly classify borrowers if the attributes it employs to evaluate the credit score are at least nearly correct. As discussed in Dejaeger *et al.* [4], the accuracy and reliability of data is of high importance, and in the cases of insufficient data quality, quantitative analysis may fail to provide the desired results, even leading to incorrect decisions being drawn. Canada's main banking regulator [34], the Office of the Superintendent of Financial Institutions, has recently suggested Canadian banks should verify borrower's income for mortgage loans. With a long history of use of the same credit scoring model, borrowers learn the characteristics being used to grant loans. Bad borrowers are motivated to falsify their attributes in order to increase their chances of getting loans. In Chapter 1, we introduced the concept of bust-out-fraud, where borrowers purposely build up and maintain a good credit history, they will then use all the available credit in one time to borrow a large amount of loan, and then abandon the account. We consider this in some sense to be like lying about the attribute since behavior is modified to get a desired score rather than for its own intrinsic rewards.

In this chapter, we will study different attitudes and behaviors of borrowers that affect loan decisions made by the banks. We model the lending business as a game performed between banks and borrowers. Borrowers can purposely report their characteristics in a way which favors loan approval. On the other hand, banks can restrict their classifier and tighten credit. We assume that both borrowers and lenders act so as to maximize their utility. Indeed, under normal circumstances, there is an equilibrium between the utility of the bank and the borrowers. At this equilibrium state, banks' strategy of maximizing expected return and minimizing losses is profitable and creates an optimal utility; on the other hand, all borrowers are treated fairly and maintained at their own favorable utility level. Borrowers with credit score higher than the cutoff score calculated by the bank will be granted credit, otherwise credit will be declined. With the advent of knowledge on the attributes being used in the scoring model, bad borrowers learn the best way to lie about their characteristics, which can increase their chances of getting loans. However, having liars in the portfolio contribute to generating negative utility and profit to the banks. We will study the stepwise reaction of the banks and borrowers in the following order:

Step 1. Present the optimal utility level of good and bad borrowers in the equilibrium state.

Step 2. Study the effect of having bad liars on the bank's optimal utility level.

Step 3. Present the primary remediation process of the bank to update the credit scoring model in order to increase its utility.

Step 4. Show that the bank can further increase its utility by exerting effort to verify the correctness of borrowers' reported attributes.

Step 5. Analyze the reaction of bad liars to maintain at their optimal utility level.

Step 6. Show that it is possible for the lending business to return to its equilibrium status.

Different statistical measures will be used to answer the questions: 1) How did the bank first notice that there are liars in the portfolio? 2) What is the effect of having bad liars in the portfolio? 3) Is it worth spending money to check for lies? If yes, 4) What is the optimal amount to spend and is it worth to spend money as a precaution in case there are liars in the portfolio?

In addition, we will examine two things: First, the cost required to make a clever lie, that is, a lie that increases the chances of loan approval while reduces the chance of detection. Next, the cost required to verify the correctness of borrowers' reported attributes. Particularly, we will look into these two costs towards the behavior of the banks and bad liars. We will also examine if the cost for bad liars to make a clever lie affects their incentive to lie and the lie intensity that they should add onto their attribute; if the cost required to verify the data of an application alters the banks' decision of spending money, and also the exact amount of money that they should distribute to check for lies. We will show that the stepwise interaction stated above will terminate and reach a steady state at one of the intermediate steps if either one of these two costs are too high. In particular, different combinations of these two costs will result in four different scenarios on the behavior of the bank and bad liars. We will present and discuss our results in each of the scenarios.

Unlike Chapter 4, we model borrowers' characteristics to come up with the granting decision. This chapter focus on modeling credit scores of borrowers to determine the cutoff score. In Section 5.3, we present evidence from a real dataset that borrowers' FICO score can be approximated by triangular distribution. We will assume credit scores of borrowers follow half triangular distribution. A thorough discussion of this assumption will be presented in Section 5.3. The half triangular distribution has been widely used in different research areas. Woo [13] considered inference on reliability, and derived the K-th moment of the ratio of two independent half triangular distributions with different support. Kim *et al*. [10] applied half triangular distribution on Bayesian estimation to derive shape parameter and computed the mean square error.

The succeeding sections are organized as follows. Section 5.2 presents the method and model used in this research. In Section 5.3, we present data to suggest that a half triangular distribution model might be reasonable. Section 5.4 presents the objective functions used to explain the interaction between the banks and bad liars. In Section 5.5, we will present and

analyze the parameters used in this research. Section 5.6 uses the data described in Section 5.3 to evaluate the objective functions mentioned in Section 5.4 to obtain our results. Conclusions are discussed in Section 5.7.

5.2 Method and Model

This chapter uses discriminant analysis as presented in Section 2.1. Notice that this research directly models credit scores of borrowers to determine the cutoff score. The method used to come up with the credit score, the characteristics of borrowers involved in the calculation of their respective credit score and the accuracy of the score on reflecting the repayment behavior of borrowers are all not the focus of this chapter, but will be suggested in future studies. We apply the set inequality in Eq. (2.5) of Section 2.1 and assume X to be the credit scores of borrowers. Our dataset contains four different types of borrowers in the lending business:

1. Good borrowers who are honest on their loan application form and will always repay their loans.

2. Good borrowers who lie on their loan application form but always repay their loans.

3. Bad borrowers who are honest on their loan application form but do not repay their loans.

4. Bad borrowers who lie on their loan application form and do not repay their loans.

We assume our dataset contains borrowers in 1, 3 and 4. In other words, we assume all good borrowers are honest and all bad borrowers (bad liars and truth tellers alike) will default. Since we build our model using historical data with known repayment performance, we assume we have precise information on whether the borrowers lied about their characteristics and also the amount of lies. We make an assumption that the size of the lie is the difference between the credit score evaluated using all correct attributes and the credit score evaluated using the reported attributes with borrowers' lies.

To study the interaction between the bank and bad liars, we further segment bad payers into truth tellers and liars. Bad truth tellers were honest on their attributes but they do not have the ability to repay their loans. They were granted loans due to inefficiencies during the credit evaluation process; or perhaps it was simply due to bad luck that they defaulted at the end. Bad liars, on the other hand, falsified their attributes, successfully escaped from credit checks to obtain loans but also defaulted. In our studies, we assume only bad payers will lie about their attribute that can increase their credit score X_B. We assume that good payers do not lie, and added noise to X_B only. Not all bad payers choose to lie about their characteristics; we

introduced another Bernoulli random variable Noi, which takes the value one when a particular bad payer lies, otherwise taking the value zero. The probability that a bad payer will lie is P_N. We further assume that all bad payers who intended to lie will result in adding a positive constant A to the correct credit score. In addition, to illustrate the effect of eliminating lies, we assume a parameter η which represents the amount of effort spent on enriching borrowers reported attributes, and reduces the magnitude of lies in credit score X_B from A to $A(1-\eta)$. Note that, in our case, η will be a value between 0 to 1. For example, a liar with credit score 600 might alter his attributes in a way which returns a credit score of 700. In this case, since the borrower lied, Noi takes the value 1 and A would be 100. If the lender did not put any effort to verify the attributes of the borrowers ($\eta = 0$), then the reported score will equal 700, if the lender put some effort ($\eta = 0.5$) on checking the attributes, then the reported score will be 650, closer to the correct value. If $\eta = 1$, that refers to the case where the lender carefully checks the attribute values reported by the borrower and successfully changed the reported value to its correct value. In real situations, it is extremely difficult for lenders to exclude all the lies that borrowers make. However, the chances of having all the lies excluded for a particular borrower is not impossible, and our model captures that. Furthermore, the incorrect value of the attributes may not be the cause of an intended lie. For instance, it is difficult to determine the value of a house: different appraisers might assign a slightly different value for the same house. Therefore, even if the lender found some discrepancies between the reported value and the truth value of the borrower's attributes, as long as the lender truly thinks the borrower has a high chance of repaying, the loan should be approved. Note that we simplified our research by assuming all discrepancies from the truth value of attributes are due to a lie. As a result, the credit score of all the bad payer is

$$\begin{aligned} X_{Bnew} &= \begin{cases} X_B, & \text{if the bad payer is a truth teller} \\ X_B + A(1-\eta), & \text{if the bad payer is a liar} \end{cases} \\ &= X_B + Noi \times A(1-\eta), \text{where} \quad Noi \sim \text{Bern}(P_N). \end{aligned} \tag{5.1}$$

[Note that X_B is independent of Noi.]

5.3 Data Description

This research directly model credit scores of borrowers to determine the cutoff score used to classify good and bad payers. In order to have an insight of the distribution of FICO scores, we

retrieve a consumer credit loan dataset in Lending Club Corporate website[2] for loans issued in 2014. Lending Club is an example of a popular social lending media company, also known as peer-to-peer lending, where lenders and borrowers can do business without the help of institutional intermediaries such as banks. One can refer to Appendix B for more information. The Lending Club dataset contains approximately 236K borrower records with credit information normally used by credit agencies to come up with the granting decisions. The FICO scores in the dataset are reported as two numbers in the form of FICO low and FICO high. Following the approach of Malekipirbazari and Aksakalli [35], we averaged the two FICO numbers and call the result the FICO score. Figure 5.1 shows the histogram of FICO scores which exhibit a positively skewed normal trend. Notice that this dataset only contains applicants who have successfully obtained their loans; as is all too typical, we have no information on borrowers who have applied but got rejected on their loan application. Therefore, no reject inference analysis can be done to predict the performance of borrowers who have requested a loan but got rejected. Furthermore, borrowers who got their loan approved normally have a higher FICO scores compared to borrowers who have been declined. Therefore, one can expect the distribution of FICO scores for all applicants to approximately follow a normal trend. Since many of the records in the dataset have not reached maturity, we filter the data to only include loans with a status of fully paid or defaulted. There are approximately 74K loans that are fully paid and 21K loans that are defaulted. Borrowers who fully paid off their loans are good payers and we found their average FICO score to be 696, while bad payers who defaulted have an average FICO score of 687, which, as expected is lower than that of good payers. Note that there is a very small difference between the average FICO score of good and bad payers, that is due to the fact that the dataset only contains successful loan applicants. If reject inference had been considered and included applicants who have their loan request rejected, the average FICO score of bad payers should be lower than 687. Therefore, it is reasonable to approximate FICO scores of good and bad payers as shown in Figure 5.2. Since granting decisions can easily be made for borrowers with FICO scores at the extreme ends (very high or very low FICO scores), classifiers should focus on borrowers with FICO scores in the middle range. Therefore, we consider using half triangular distribution to approximate credit scores of borrowers, which captures the middle range and ignores the two tails at the extreme ends.

The rest of the sections are organized as follows. Section 5.3.1 presents an overview and justifies the usage of the triangular distribution. In Section 5.3.2, we describe the credit scores of good payers and derive the probability density function using half triangular distribution. Section 5.3.3 describes and presents the probability density function of credit scores of bad truth tellers and bad liars.

[2] `http:\\www.lendingclub.com.`

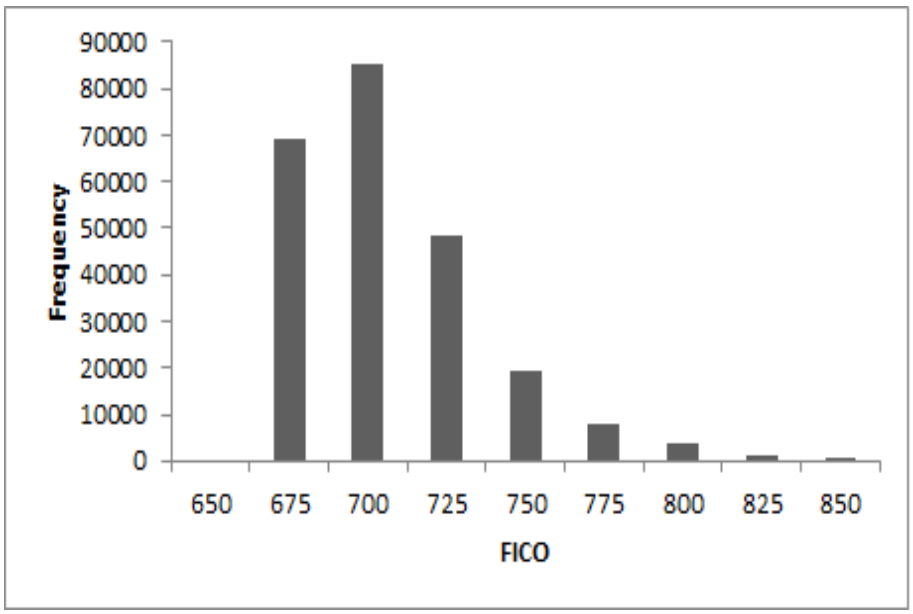


Figure 5.1: The dataset in Lending Club website contains 236K borrowers' records and reported FICO scores as two numbers: FICO low and FICO high. FICO scores of borrowers have been calculated as the average of FICO low and FICO high. This histogram exhibit FICO scores with a positively skewed normal trend.

5.3.1 Overview of Triangular Distribution

This research uses a triangular distribution to illustrate the game between the banks and borrowers in the lending industry. The earliest mention of the triangular distribution dates back to Simpson [15] in 1755. Since then, the study of the triangular distribution seems to have vanished in the statistical literature, until R. Schmidt [16] presented the probability density function of symmetric triangular distribution in 1934. In 1941, Ayyangar [17] presented a brief study of non-symmetric standard triangular distribution. In the past years, due to the over simplicity of the modeling assumptions towards the application in real data, researchers somehow neglected the statistical importance of triangular distribution. Recently, the benefit of using a simpler distribution on research work provides a more obvious understanding on the analysis of results, in addition, the ease of estimating parameters of the triangular distribution contributed to more work on financial applications being done using this distribution. This motivates Johnson and Kotz [18] to revisit and provide more details of the triangular distribution. In particular, Johnson [19] uses it as a proxy for beta distribution in risk analysis to leverage the difficulty on parameter estimation. Different uses of the triangular distribution correspond to different number of parameters and symmetry of the triangle can be found in past research works. Research conducted by Clark [21], Grubbs [22], MacCrimmon and Ryaveck [23], Moder and Rodgers [24], Vaduva [25], Williams [26], Keefer and Bodily [27] and D. Johnson [20] all use

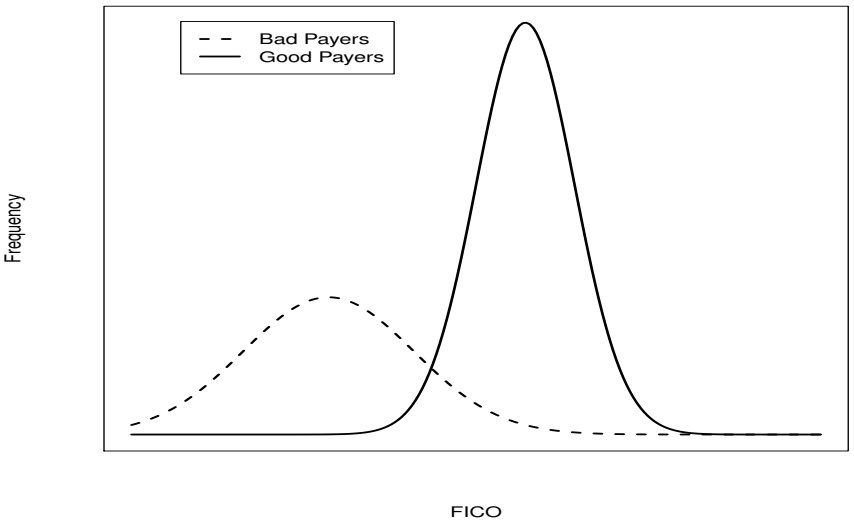


Figure 5.2: Approximate FICO trend of good and bad payers.

the asymmetric three-parameter triangular distribution presented in Figure 5.3. This asymmetric triangle has base $d-c$ and height $\frac{2}{d-c}$. The probability density function of this asymmetric three-parameter triangular distribution is presented in Eq. (5.2).

$$f(x) = \begin{cases} \frac{2}{d-c}\frac{x-c}{m-c}, & \text{if } c \le x \le m \\ \frac{2}{d-c}\frac{d-x}{d-m}, & \text{if } m \le x \le d \\ 0, & \text{otherwise.} \end{cases} \tag{5.2}$$

Our research work on the stepwise change of the utility of different borrowers will be studied by applying a simpler version of the triangular distribution stated in Eq. (5.2); the half triangular distribution. By setting either the value of c or d to be equal to m, the triangle in Figure 5.3 becomes either a right or left half triangle respectively. As described at the beginning of this section, credit scoring models should focus on classifying borrowers with their credit scores in the middle range. By ignoring the extreme ends; left tail of bad borrowers and right tail of good borrowers, it is reasonable to use half triangular distribution as presented in Figure 5.4 to do our research.

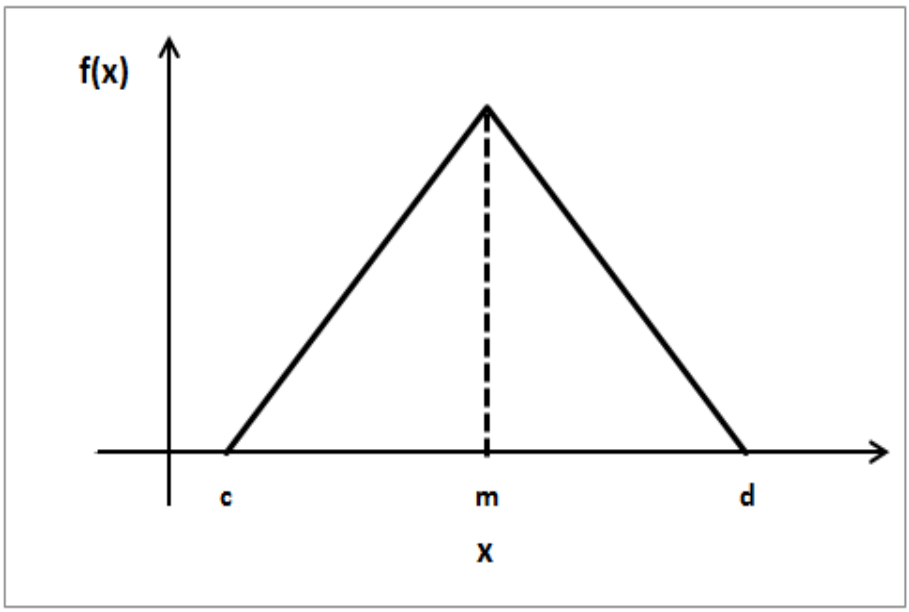


Figure 5.3: The probability density function of a asymmetric triangular distribution with support [c,d], mode m and height $\frac{2}{d-c}$.

5.3.2 Credit scores of good borrowers

We will focus on classifying good borrowers with a relatively low credit score, therefore, we assume credit scores of good borrowers follow a left triangular distribution and with values varying from $2a$ to $4a$. In other words, we set c to $2a$, and d to m to $4a$ in Figure 5.3. Notice that left triangular distribution has higher density on larger values. The probability density function of good borrowers can be found in Eq. (5.3). The solid lines in Figure 5.4 plot the probability density function of the credit scores of various classes of borrowers, including good borrowers.

$$f(x|G) = \begin{cases} 0, & \text{if } x < 2a \\ \frac{x-2a}{2a^2}, & \text{if } 2a \le x \le 4a \\ 0, & \text{if } x > 4a \end{cases} \tag{5.3}$$

5.3.3 Credit scores of bad borrowers

As discussed in Section 5.2, bad borrowers can be further subdivided into the group of bad truth tellers and bad liars. If bad liars did not lie, their correct score should have the same distribution as the bad truth tellers. In this research, the effect of adding a lie of size A and

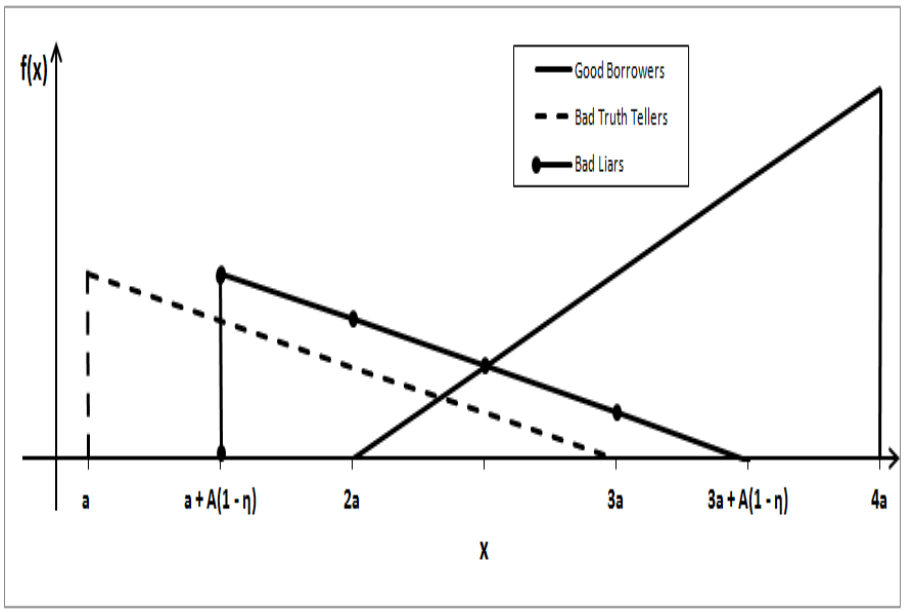


Figure 5.4: This plot presents the probability density function (PDF) of the credit scores of different borrowers. The solid line displays the PDF of the scores of good borrowers, which follows left triangular distribution and has values varies from $2a$ to $4a$. The dash line shows the PDF of bad truth tellers, which follows right triangular distribution and has value varies from a to $3a$. The solid line with circle pattern shows the PDF of the credit scores of bad liars. It follows right triangular distribution. A constant amount of A represents the amount of lies that bad liars altered their attributes and resulted in an increase on their score. A parameter η represents the amount of effort spend on eliminating lies. The mix effect of A and η causes the credit scores of bad liars to vary from $a + A(1-\eta)$ to $3a + A(1-\eta)$.

spending η amount of effort to correct the score, shifted the characteristic of bad liars from X_B to $X_B + A(1-\eta)$ as shown in Eq. (5.1).

Bad truth tellers

Again, we want to focus on classifying bad borrowers with relatively high credit scores, therefore, we assume the credit scores of bad truth tellers to follow right triangular distribution. That is to say, we set c equal to m equal to a, and d equal to $3a$ in Figure 5.3. Notice that right triangular distribution has higher density at its low values. The probability density function of bad truth tellers are presented in Eq. (5.4). The dash line in Figure 5.4 plotted the probability density function of the scores of bad truth tellers.

$$f(x|B) = \begin{cases} 0, & \text{if } x < a \\ \frac{3a-x}{2a^2}, & \text{if } a \le x \le 3a \\ 0, & \text{if } x > 3a \end{cases} \tag{5.4}$$

Bad liars

As discussed, the distribution of the scores of bad truth tellers and liars should be the same. Therefore, the scores of bad liars also follows right triangular distribution. Since the score of bad truth tellers varies from a to $3a$, applying Eq. (5.1), the score of bad liars varies from $a+A(1-\eta)$ to $3a+A(1-\eta)$. Note that bad liars have scores larger than bad truth tellers, which makes them more likely to receive loans. The probability density function of bad liars can be found in Eq. (5.5). The solid line with circle pattern in Figure 5.4 plotted the probability density function of the scores of bad liars.

$$f(x|Blied) = \begin{cases} 0, & \text{if } x < a + A(1-\eta) \\ \frac{3a+A(1-\eta)-x}{2a^2}, & \text{if } a + A(1-\eta) \le x \le 3a + A(1-\eta) \\ 0, & \text{if } x > 3a + A(1-\eta) \end{cases} \tag{5.5}$$

Applying the relationship between the scores of bad truth tellers and liars in Eq. (5.1), the probability density function of all the bad payers is:

$$\begin{aligned} f(x|Bnew) &= \begin{cases} 0, & \text{if } x < a \\ (1-P_N)\frac{3a-x}{2a^2}, & \text{if } a \le x < a + A(1-\eta) \\ (1-P_N)\frac{3a-x}{2a^2} + P_N\frac{3a+A(1-\eta)-x}{2a^2}, & \text{if } a + A(1-\eta) \le x < 3a \\ P_N\frac{3a+A(1-\eta)-x}{2a^2}, & \text{if } 3a \le x < 3a + A(1-\eta) \\ 0, & \text{if } x \ge 3a + A(1-\eta) \end{cases} \\ &= \begin{cases} 0, & \text{if } x < a \\ (1-P_N)\frac{3a-x}{2a^2}, & \text{if } a \le x \le a + A(1-\eta) \\ \frac{3a-x+P_N A(1-\eta)}{2a^2}, & \text{if } a + A(1-\eta) \le x \le 3a \\ P_N\frac{3a+A(1-\eta)-x}{2a^2}, & \text{if } 3a \le x \le 3a + A(1-\eta) \\ 0, & \text{if } x > 3a + A(1-\eta) \end{cases} \end{aligned} \tag{5.6}$$

Referring to Figure 5.4 and Eq. (5.6), if the reported score is between a and $a+A(1-\eta)$, we are certain that the applicant is a bad truth teller. However, if we know that the reported score is from a bad borrower and is anywhere between $a+A(1-\eta)$ to $3a$, the applicant maybe a bad true teller or a bad liar, with probability P_N being a bad liar. Furthermore, if we know that the reported score is from a bad borrower and has value greater than $3a$, the borrower is a bad liar.

5.4 Modeling Borrower and Lender Objective Functions

This section presents the objective functions used to quantify the behavior first of the bank and next of the bad liars. We will evaluate the value of all these functions in each of the six steps discussed in the introduction to explain the interactive game performed between the bank and bad liars. We fix the number of loan applicants to be N in each step. In addition, we will evaluate the objective functions for good borrowers and bad truth tellers to study how they have been affected consequently. Notice that our research studies the amount of lies A that bad liars added onto their attributes to positively affect their credit score and the corresponding countervailing effort η that banks put to eliminate lies. All the objective functions discuss in this section will depend on both A and η. Details of these functions are provided below:

5.4.1 The cutoff score

The set inequality stated in Eq. (2.5) of Section 2.1 presents the credit scoring model used in this research, and we denote $\gamma(\eta, A)$ to be the cutoff score of the scoring model. As in previous chapters, we restrict our work on loans with a repayment of principal and interests, the retail equivalent of a bullet bond [36]. Borrowers need not pledge any collateral on their loans, and in case of a default, the bank will not be able to obtain any recovery (that is, the loss given default is equal to 100%, in line with the guidelines for foundational Basel II credit risk systems [30]). We denote L to be the amount of loan requested, assumed the same across all borrowers, and will be charged with the same amount of interest i. Thus $L \times i$ is the total amount of interest paid. In other words, the expected misclassification profit Q that the lender might have been able to earn is lost if a good payer was classified as a bad payer is $L \times i$. Furthermore, there are only good and bad payers among all the borrowers, it implies that $p_B = (1 - p_G)$. Our research captures two types of bad payers: bad truth tellers and bad liars. We derive the probability density function of all the bad payers in Eq. (5.6) and denoted it as $f(x|Bnew)$. Putting in all these conditions, the set inequality in Eq. (2.5) of Section 2.1 becomes

$$A_G = \left\{ x \middle| \frac{L(1-p_G)}{Lip_G} \leq \frac{f(x|G)}{f(x|Bnew)} \right\}. \tag{5.7}$$

Note that the above set inequality gives the cutoff score $\gamma(\eta, A)$ of our research. We will refer to Figure 5.4 to determine the possible range for $\gamma(\eta, A)$. If $\gamma(\eta, A)$ is less than $2a$, all the good borrowers will be granted loans, violating the idea of using discriminant analysis to maximize expected profit while minimizing expected loss, since if $\gamma(\eta, A)$ changes, the expected profit will remain constant. On the other hand, if $\gamma(\eta, A)$ is greater than $3a$, none of the bad truth tellers will be granted loans, it will be more realistic to include a portion of bad truth tellers

who will have their loans approved but defaulted at the end. Therefore, we decided to restrict the cutoff score to be between $2a$ and $3a$ throughout the paper. By substituting Eq. (5.3) and (5.6) into the R.H.S. of the set inequality of Eq. (5.7), it becomes

$$\frac{f(x|G)}{f(x|Bnew)} = \frac{\frac{x-2a}{2a^2}}{\frac{3a-x+P_NA(1-\eta)}{2a^2}} = \frac{x-2a}{3a-x+P_NA(1-\eta)}. \tag{5.8}$$

Putting the above back into the set inequality in Eq. (5.7), it gives

$$\begin{aligned} A_G &= \left\{ x \middle| \frac{L(1-p_G)}{Lip_G} \leq \frac{\frac{x-2a}{2a^2}}{\frac{3a-x+P_NA(1-\eta)}{2a^2}} \right\} \\ &= \left\{ x \middle| \frac{(1-p_G)}{ip_G} \leq \frac{x-2a}{3a-x+P_NA(1-\eta)} \right\} \\ &= \left\{ x \middle| x \geq \frac{3a(1-p_G)+P_N(1-\eta)A(1-p_G)+2aip_G}{ip_G+(1-p_G)} \right\}. \end{aligned} \tag{5.9}$$

From the above set inequality, we know that the cutoff score of our scoring model is

$$\gamma(\eta, A) = \frac{3a(1-p_G)+P_N(1-\eta)A(1-p_G)+2aip_G}{ip_G+(1-p_G)}. \tag{5.10}$$

5.4.2 The number of loans granted

We will capture the number of loans granted and use this as a reference to detect any abnormal behavior among different borrowers. As mentioned in Section 5.4, we assumed there are N loans granted at each of the six steps that we will study, and if there is no change in any lending policy to tighten or loosen credit, the total amount of loans granted should be similar in each step. In real situations, banks should have a similar amount of ending net receivables and origination volumes each month. Since we assumed the loan size and interest rate to be the same amongst all borrowers, looking at the total number of loans issued is the same as looking at the total dollar amount granted.

When bad liars change their correct attribute, there will be more bad payers with scores greater than the cutoff value. This will result in more loans granted to bad payers and thus, there will be an increase in the total number of loans granted. This acts as the first warning signal for the banks to be aware that something might have gone wrong. We denote N_G, N_B and N_{Blied} to be the number of loans granted to good borrowers, bad truth tellers and bad liars respectively. Notice that the number of loans granted to these borrowers is the number of borrowers with their corresponding score greater than the cutoff score. Furthermore, we denote N_T to be the total number of loans issued. It follows that N_T is the sum of N_G, N_B and N_{Blied}.

Applying the probability density functions in Eq. (5.3), (5.4) and (5.5), we know that

$$\begin{aligned} N_G &= Np_G \int_{\gamma(\eta,A)}^{\infty} f(x|G)\,dx \\ &= Np_G \int_{\gamma(\eta,A)}^{4a} \frac{x-2a}{2a^2}\,dx \\ &= Np_G \left[\frac{\gamma(\eta,A)}{a} - \frac{\gamma^2(\eta,A)}{4a^2}\right], \end{aligned} \tag{5.11}$$

$$\begin{aligned} N_B &= N(1-p_G)(1-P_N) \int_{\gamma(\eta,A)}^{\infty} f(x|B)\,dx \\ &= N(1-p_G)(1-P_N) \int_{\gamma(\eta,A)}^{3a} \frac{3a-x}{2a^2}\,dx \\ &= N(1-p_G)(1-P_N) \left[\frac{9}{4} - \frac{3\gamma(\eta,A)}{2a} + \frac{\gamma^2(\eta,A)}{4a^2}\right], \end{aligned} \tag{5.12}$$

$$\begin{aligned} N_{Blied} &= N(1-p_G)P_N \int_{\gamma(\eta,A)}^{\infty} f(x|Blied)\,dx \\ &= N(1-p_G)P_N \int_{\gamma(\eta,A)}^{3a+A(1-\eta)} \frac{3a+A(1-\eta)-x}{2a^2}\,dx \\ &= N(1-p_G)P_N \left[\frac{(3a+A(1-\eta))^2}{4a^2} - \frac{(3a+A(1-\eta))\gamma(\eta,A)}{2a^2} + \frac{\gamma^2(\eta,A)}{4a^2}\right] \\ &= N(1-p_G)P_N \left[\frac{(3a+A(1-\eta)-\gamma(\eta,A))^2}{4a^2}\right] \end{aligned} \tag{5.13}$$

and finally

$$N_T = N_G + N_B + N_{Blied}. \tag{5.14}$$

5.4.3 The utility function of different parties

The interaction of the bank and bad liars affects the utility level of the different parties. Note that higher utility value represents more satisfaction of each party. Bad liars increase their utility level by adding a quantity A of lies to their score, however, this will lower the utility level of the banks. The bank can then choose either to directly catch liars or to tighten credit in the portfolio to remediate the amount of losses incurred, the results of those two actions will increase the utility of banks, but may affect the utility of other parties in different directions.

The following shows the derivation of the utility function of different parties.

For the banks

We define the utility function of the bank to be its profit per loan, which is the revenue earned less the cost spent on giving out loans. We denote $E[Rev(\eta, A)]$ to be the total expected revenue of the bank. In our assumption, the bank earns only the interest amount of $L \times i$ from each good payers, and loss the loan amount of L in the event of a default from bad payers. Therefore, the expected revenue of the bank is the amount of interest earned from each good payer multiplied by the number of loans granted to good payers, less the loss loan amount of L from each defaulted bad payer multiplied by the number of loans granted to bad payers. Notice that the number of loans granted to bad payers is the total of the number of loans granted to bad truth tellers and bad liars. Therefore,

$$
\begin{aligned}
&E[Rev(\eta, A)] \\
&= LiN_G - L[N_B + N_{Blied}] \\
&= LiN_G - LN_B - LN_{Blied} \\
&= LiNp_G\left[\frac{\gamma(\eta,A)}{a} - \frac{\gamma^2(\eta,A)}{4a^2}\right] \\
&- LN(1-p_G)\left[\frac{9}{4} - \frac{(3a + P_N A(1-\eta))\gamma(\eta,A)}{2a^2} + \frac{\gamma^2(\eta,A)}{4a^2} + \frac{(3a + A(1-\eta))^2 P_N}{4a^2} - \frac{9P_N}{4}\right], \quad (5.15)
\end{aligned}
$$

where the last equality holds if we substitute Eq. (5.11), (5.12) and (5.13) for N_G, N_B and N_{Blied} respectively. Notice that if the bank can correctly classify all the applicants, the maximum potential gain is Np_GiL, and the maximum potential loss is $N(1-p_G)L$.

Loan application costs involved in processing all applications include credit checks, property appraisals for mortgage loans or properties pledged as collateral, and basic administrative costs [2]. Thus, it is clear that more effort spent on eliminating lies will result in higher costs, and this cost will apply to all loan applicants. To simplify our model, we fix the unit cost of checking lies and assume the magnitude of lies does not affect the cost of eliminating them. We further assumed that the total cost spent on granting loans and eliminating lies to be linear in η. We define the total cost to be $Nk\eta$, where k is the proportionality constant which represents the unit cost of effort spent to eliminate lies. Therefore, the utility function of the bank is

$$U_{BK}(\eta, A)$$

$$
\begin{aligned}
&= E[Rev(\eta, A)] - N k\eta \\
&= L i N_G - L N_B - L N_{Blied} - N k\eta \\
&= L i N p_G \left[\frac{\gamma(\eta, A)}{a} - \frac{\gamma^2(\eta, A)}{4a^2} \right] \\
&- L N (1 - p_G) \left[\frac{9}{4} - \frac{(3a + P_N A(1-\eta))\gamma(\eta, A)}{2a^2} + \frac{\gamma^2(\eta, A)}{4a^2} + \frac{(3a + A(1-\eta))^2 P_N}{4a^2} - \frac{9P_N}{4} \right] - N k\eta,
\end{aligned}
\tag{5.16}
$$

where the last equality follows if we substitute Eq. (5.15) for $E[Rev(\eta, A)]$.

For the borrowers

As described by Anderson [8], credit allows borrowers to buy now and pay later. Borrowers who were granted loans used the borrowed money to purchase goods and services to fulfill their needs. Roberts and Jones [31] studied money attitudes of college students and referred to money as a tool of power. Csikszentmihalyi and Rochberg-Halton [32] described money as a form of power that consists of respect, consideration and envy from others and represents the goals of a culture. Borrowers will obtain certain level of satisfaction through the use of the money. Easterlin [37] stated that more money typically means more happiness. The utility of borrowers should be correlated to the amount of loan that was granted. Furthermore, Fernald *et al*. [38] studied the psychological effects on granting microcredit loans to poverty groups in developing countries and concluded that individuals who were granted credit exhibit lower levels of depressive symptoms. In other words, borrowers who applied but did not receive a loan, either due to the misclassification of the credit scoring model or because they truly do not have the ability to repay, should have negative utility. It is very difficult to measure the dissatisfaction of borrowers who have their loan application rejected, therefore, to simplify our research work, we assume borrowers who applied for the loan but did not get it granted have zero utility.

Many, although not all, good borrowers were granted loans and in return must repay the loan amount and interest to the bank. Denote I to be the percentage of benefit that good borrowers gained with the use of the loan, where I is greater than the interest rate i, such that the satisfaction that good borrowers obtained from using the loan is greater than the amount of interest that they have to repay. Therefore, the amount of benefit that each good borrower gains is $L(I - i)$, and the total amount of utility for all the good borrowers is

$$
U_G(\eta, A) = L(I-i)N_G = LIN_G - LiN_G = L(I-i)N p_G \left[\frac{\gamma(\eta, A)}{a} - \frac{\gamma^2(\eta, A)}{4a^2} \right], \tag{5.17}
$$

where the last equality follows if we substitute Eq. (5.11) for N_G. We need to model the utility of good borrowers. Honest good borrowers perceive themselves to be better off with the loan then without it; in other words they receive a utility of $L(1+I)$ by receiving a loan L and must repay $L(1+i)$ for a relative utility of $L(I-i)$. Later we will change I to 40% and i to 30% to demonstrate the results.

Bad truth tellers were honest about their attributes and truly do not have the ability to repay their loan. They were granted loan due to the misclassification of the credit scoring model. After they received their loan, they will use it to satisfy their needs. We assume the utility of each bad truth teller equals to the amount of loan that they received. Therefore, the total utility for all the bad truth tellers is

$$U_B(\eta, A) = LN_B = LN(1-p_G)(1-P_N)\left[\frac{9}{4} - \frac{3\gamma(\eta,A)}{2a} + \frac{\gamma^2(\eta,A)}{4a^2}\right], \tag{5.18}$$

where the last equality follows if we substitute Eq. (5.12) for N_B.

Bad liars exert effort to learn about the attributes used to make granting decisions in the credit scoring model. Afterward, they have to figure out a safe way to lie effectively and without being caught. In reality credit fraud, once detected, can cause very serious harm to a borrower's credit history. The liars name might be put onto a fraudulent list under all credit bureaus system, such that all financial institutes who obtain credit information from the bureaus before finalizing their granting decision will know about their previous fraudulent behavior and likely decline their application. Equifax Canada Inc. launched a fraud prevention and management tool in 2010; Citadel [33], to match new applications with a universe of previous applications to uncover suspicious behavioral patterns of data that suggest fraudulent activity. Thus, it is very important for bad liars to make a clever lie, and under our assumptions, a larger amount of lie added to the correct attribute corresponds to higher risk. For instance, it is possible for a waitress to earn $500 extra per month as tips in a busy high-end restaurant, however, it is suspicious if the waitress reported $5000 as the extra amount of tips that she earned each month. Therefore, it is reasonable to assume that the amount of lie added to the correct attribute is linear to its cost. We define the total cost of lying to be Ah, where h is the proportionality constant. Note that this cost applied to all the bad liars, no matter whether they received their loans or not. We further fix the unit cost of lying to be the same amongst all borrowers and assume that all borrowers lie at the same level. Therefore, the total utility of bad liars is

$$U_{Blied}(\eta, A) = LN_{Blied} - (1-p_G)P_N NAh$$

$$= LN(1-p_G)P_N\left[\frac{(3a+A(1-\eta)-\gamma(\eta,A))^2}{4a^2}\right]-(1-p_G)P_N NAh, \tag{5.19}$$

where the last equality follows if we substitute Eq. (5.13) for N_{Blied}. In general the overall utility for all the borrowers is

$$\begin{aligned} U_T(\eta,A) &= U_{BK}(\eta,A)+U_G(\eta,A)+U_B(\eta,A)+U_{Blied}(\eta,A) \\ &= LiN_G - LN_B - LN_{Blied} - Nk\eta + LIN_G - LiN_G + LN_B + LN_{Blied} - (1-p_G)P_N NAh \\ &= LIN_G - Nk\eta - (1-p_G)P_N NAh, \end{aligned} \tag{5.20}$$

where the second equality holds if we substitute Eq. (5.16), (5.17), (5.18) and (5.19) for $U_{BK}(\eta,A), U_G(\eta,A), U_B(\eta,A)$ and $U_{Blied}(\eta,A)$. The last equality stated the overall utility of the lending business, it equals to the total amount of interest earned from good borrowers less the total amount banks spend on checking lies and the total amount bad liars spend to make lies.

5.5 Parameters and Analysis

We focus on lending businesses that target customers in the sub-prime category. Borrowers with higher credit risk [28] may choose to borrow loans from sub-prime lending institutions, which demand higher interest rates, but in return require fewer credit checks and which provide easier and faster access to cash. We will sub-divide this section into two parts. In the first part, we will fix some of the parameters used in this research to correspond to the conditions of a sub-prime lending institution. In the second part, we will analyze the unit cost to lie h and the unit cost to check for lies k. In particular, we will show that there are four different combinations of h and k which translate into four different scenarios.

5.5.1 Fixed Parameters

The fixed parameters used for this research are shown in Table 5.1. We fix the number of loan applicants N to be 1MM. The loan amount L to be \$1K and is the same across all borrowers. Therefore, the entire portfolio consist of \$1BN exclusive of interest charges[3]. We fix the interest rate i to be at 30% per annum. Credit card and department store card issuers typically range their interest rate from 7% to 36%, depending upon the bank's risk evaluation methods and borrowers' credit history. As mentioned, we assume good borrowers gain benefit of 40% with the use of the loan. In other words, good borrowers can gain \$400 worth of benefit but

[3]Only principal is counted.

Table 5.1: Parameters used to generate results for each of the six steps described in the introduction .

N	L	i	I	p_G	P_N	a
1MM	\$1K	30%	40%	80%	50%	200

have to repay \$300 as interest, therefore, the net gain of a good borrower is \$100, which is 10% of the amount borrowed. On the other hand, since interest rate has been fixed at 30%, banks gain \$300 if the customer is a good payer and lose \$1K if the customer is a bad payer. The Lending Club dataset discussed in Section 5.3 had 74K of good payers and 21K of bad payers, so roughly 80% good payers. Therefore, we set p_G to 80%; meaning that 80% of the loan applicants are good payers, while 20% of them are bad payers. That is to say, if the bank can perfectly classify all the applicants, the potential gain of the bank is 1MM×30%×\$1K ×80% = \$240MM which is greater than the potential loss of 1MM×\$1K×20% = \$200MM. If the bank granted loans to all applicants without the use of any classifier, the expected revenue gain in this portfolio is 4%[4]. Since this research study the effect of borrowers lie to obtain loans, we further assume half the bad payers added lies onto their attributes in order to increase their chances of loan approval, therefore, we equate P_N to 0.5. In reality, there are different ways to examine whether borrowers lie to obtain loans from the historical dataset. For example, it will be subspicious for an 18 year old person to have an income of \$1MM per annum, in this case, this person maybe a fraud. Research on distinguishing borrowers into bad liars and truth tellers are mostly under development. Other applications on detecting liars include Walczyk *et al.* [39], who uses employee's response time and facial expressions to distinguish whether the employee lied, similar techniques can be modify and apply in the context of credit scoring.

We (arbitrarily) assume the synthetic credit score in this chapter has value ranges from 200 to 800[5]. Referring to Figure 5.4, we set a to 200, and thus the scores of bad liars vary from $200+A(1-\eta)$ to $600+A(1-\eta)$. As η can take values between 0 to 1, this equates the maximum score for bad liars to $600+A$. Together with the restriction that the maximum score has to be less than or equal to 800, this implies that A must take values between 0 and 200.

5.5.2 Modeling Cost of Lying and Cost of Checking

This research studies the interaction of bad borrowers lie and the corresponding amount of money that banks spend to validate the reported data. The unit cost of lying h can motivate bad borrowers to lie and affect the lie intensity that bad liars added to their reported attributes

[4]Expected revenue of the bank if loans were granted to all applicants: $\frac{240-200}{1000} \times 100\% = 4\%$

[5]We chose this (200,800) range to make this synthetic credit score to be simillar to FICO score.

which in turns affects their calculated credit score. Bad liars may have more incentive to lie if h is low and may even choose not to lie if h is too high. On the other hand, the unit cost of checking lies k affects the amount of effort that banks spend on verifying the correctness of the reported attributes provided by applicants which can improve the accuracy of credit scores. If k is small, banks may choose to put more effort to check for lies, and in cases where k is very large, banks may even choose not to exert effort to check for lies. In this subsection, we study the way in which the optimal behavior of bad liars to lie and of banks in checking for lies depends on different values of h and k. We assume that h and k are both independent of the lie magnitude and are identical across all loan applicants. We apply the fixed parameters stated in Table 5.1 and arbitrarily set the range of h to vary from \$0 to \$3 and the range of k to vary from \$0 to \$70, both with step size \$0.01. We conclude that different combinations of h and k will result in the following four different scenarios in the interactive game between the bank and bad liars:

1. Not economical to lie
2. Economical to lie but
    (a) Not economical to check for lies
    (b) Always economical to check for lies
    (c) Periodically check for lies

Figure 5.5 shows the plot of h versus k with each of the shaded regions representing one of the scenarios mentioned above. The top dark gray region has h ranging from \$2.62 to \$3 and k can be any value between \$0 to \$70. It represents the first scenario where it is not economical to lie and there will be no liars in the lending business. The light gray shaded region on the right has h ranging from \$0 to \$2.61 and k ranging from \$65.9 to \$70, represents the second scenario where it is economical to lie but not economical to check for lies. The black region also has h ranging from \$0 to \$2.61 and k ranging from \$0 to \$65.9, represents the third scenario where liars should always lie and banks should always spend money to check for lies. The white region has h ranging from \$0 to \$2.04 and k from \$27.3 to \$65.9, represents the fourth scenario where bad liars and banks should periodically lie and check for lies respectively. Note that our analysis limit the unit cost to check for lies to \$3 and the unit cost to lie to \$60, if we set h and k to some higher values, we believe Scenario 1 and 2a will have their h and k values extended to infinity respectively. We will discuss and analyze each scenario in detail in the next section.

Note that we study the interactive game between banks and bad liars with liars act first to choose whether to lie and the lie intensity, banks will then respond to sustain their profitability.

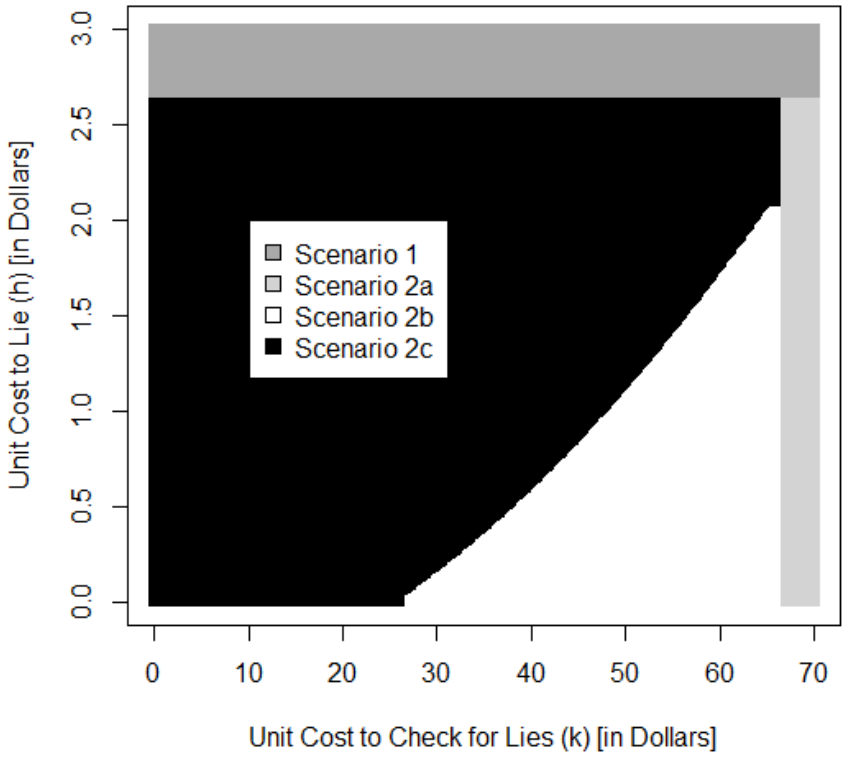


Figure 5.5: The plot above shows that different combinations of h and k results in four different scenarios, each represents a different interaction between the bank and bad liars. We varied h from \$0 to \$3 and k from \$0 to \$70 each with step size 0.1. The top dark gray region represents the scenario where it is not economical to lie. The light gray region represents the scenario where it is economical to lie but not to check for lies. The black and white regions both represents the scenario where it is economical to lie and to check for lies, with the black region representing lying and checking should be done periodically, and white region represents lying and checking should always be carried out.

The strategy of the behavior appears to depend on the relative cost of the banks checking lies and the cost of the liars to lie. When this relative cost is tilted in favor of the banks (cheap for banks to check, expensive for liars to lie) liars sometimes "lose" the game by stopping their lies, at least at some game cycles. When the relative cost is tilted in favor of the liars (cheap for liars to lie, relatively expensive for banks to check), the liars will certainly lie, and sometimes even the banks will stop checking. From bad liars' point of view, when it is cheap that banks can afford to check for lies, Figure 5.5 shows that bad liars will lie periodically and play a game with the banks on the time period that they will lie. When it is relatively more expensive to check for lies, Figure 5.5 shows that bad liars will always lie, since they know that it costs

more for banks to catch them, and if $k > \$65.9$, banks cannot even afford to spend any effort to eliminate lies. On the other hand, when the unit cost of lying is too high ($h > \$2.61$), bad liars stop lying as they cannot afford to lie. When the unit cost of lying is less than \$2.61, bad liars will lie periodically if it is cheap for banks to catch them and will always lie if it is expensive for banks to eliminate lies.

From the banks' point of view, if it is very expensive to verify borrowers' attributes, they cannot afford to check for lies and won't put effort on eliminating lies. If it is relatively cheap to validate borrowers' attributes, bad liars will always lie and banks will have to always exert effort to check for lies. When it is cheap to check borrowers' attributes for lies, bad liars will lie and banks will check for lies periodically. If it is cheap for liars to lie but relatively expensive for banks to check for lies, bad liars will always lie and banks will always check for lies. When lying is very expensive, bad liars won't be able to afford to lie, and banks do not have to spend any effort to validate borrowers' reported attributes. It is interesting to see that if the unit cost to check for lies is low, our model suggests banks to check for lies periodically. When the cost to check for lies increases, our model suggests banks to always exert effort to check borrowers' reported attributes. Since the interactive game presented here studies bad liars lie and then banks react to maintain its profitability, we have to look at the action of banks and bad liars at the same time.

5.6 Results and Analysis

This section will be sub-divided into four subsections, each describing one of the scenarios mentioned in Section 5.5.2. In each subsection, we will apply the parameters presented in Table 5.1 and arbitrarily set a value for h and k, which is within the ranges in the scenarios as described in Section 5.5.2 to imitate the interaction between the bank and bad liars. Table 5.2 provides the descriptions of the objective functions that will be presented in each of the six steps of interaction between the bank and bad liars as mentioned in the introduction. In particular, we will show that some scenarios produce fewer steps, indeed, the stepwise interaction will converge to a steady state before reaching the last step. Thorough discussion of the results and the impact of the utility of different parties will be provided.

5.6.1 Scenario 1: Not Economical to Lie

This scenario represents the situation where liars cannot afford to make lies and results in having no liars in the lending business. Using the parameters presented in Table 5.1, we generated the plot in Figure 5.5, and realized that it is not profitable for bad liars to lie if the unit cost of

Notation	Descriptions
η	Bank's effort to check for lies
A	Amount of lies that bad lies added
γ	The cutoff score
N_G	Number of good borrowers who were granted loans
N_B	Number of bad truth tellers who were granted loans
N_{Blied}	Number of bad liars who were granted loans
N_T	Total number of loans granted
LiN_G	Expected interest amount earned
LN_B	Expected loan amount loss to bad truth tellers
LN_{Blied}	Expected loan amount loss to bad liars
$Nk\eta$	The cost for banks to check for lies
U_{BK}	The utility of the bank
LIN_G	Total benefits of good borrowers
LIN_G	Total cost of good borrowers
U_G	The utility of good borrowers
U_B	The utility of bad truth tellers
LN_{Blied}	Expected loan amount granted to bad liars
$(1-p_G)P_N NAh$	The total cost for bad liars to lie
U_{Blied}	The utility of bad liars
U_T	Total utility of all parties

Table 5.2: Description of the objective functions that will be presented in the four scenarios mentioned in Section 5.5.2.

lying is greater than \$2.61 [we did not compute results for $h >$ \$3]. To illustrate this scenario, we apply the parameters presented in Table 5.1 and arbitrarily set h to \$3, calculated all the objective functions as stated in Table 5.2 and presented the results in Table 5.3. Under normal circumstances, there are no liars in the lending business, and banks do not have to put any effort to check for lies in the reported attributes of applicants. Therefore, in Step 1 presented in Table 5.3, both η and A are set to be 0. Another point of view is that there are always potential liars in the lending business, but in this scenario bad liars have not yet learned how to produce clever lies and so choose not to lie, therefore, in the results presented in Table 5.3, we separate the objective functions for bad payers into bad honest payers and bad liars independently.

Nevertheless, this results in having a cutoff score of $\gamma(\eta = 0, A = 0) = 491$, meaning that if the credit score of the applicant is greater than 491, they will be granted loans, otherwise declined. We substituted $a = 200$, $A = 0$, $\eta = 0$ and added the cutoff score $\gamma = 491$ into Figure 5.3 to generate Figure 5.6. Notice that the triangle presented with dash lines is overlapping with the triangle presented in solid lines with circle patterns, they represent the PDFs of bad truth tellers and bad liars respectively, both triangles have credit scores ranging from 200 to

	Step 1
η	0
A	0
γ	491
N_G	759K
N_B	7.44K
N_{Blied}	7.44K
N_T	774K
LiN_G	\$228MM
LN_B	\$-7.44MM
LN_{Blied}	\$-7.44MM
$Nk\eta$	\$0
U_{BK}	213MM
LIN_G	\$303MM
LiN_G	\$-228MM
U_G	76MM
U_B	7.44MM
LN_{Blied}	\$7.44MM
$(1-p_G)P_N NAh$	\$0
U_{Blied}	7.44MM
U_T	303MM

Table 5.3: Table of results for Scenario 1 where it is not economical to lie.

600. The triangle with solid lines has scores ranging from 400 to 800 representing the PDF of good borrowers. The vertical line with $x = 491$ represents the cutoff score. The gray shaded region showed the density of the scores where the corresponding applicants will be granted loans. In our assumption of having 1MM applicants, the methodology applied in this paper resulted in granting 774K loans in total, of which 759K are granted to good payers and 14.9K are granted to bad payers. Then the booking rate using this methodology is 77.4%[6]. The bank's revenue earned from good payers is \$232MM[7] and incurred losses of \$14.9MM from bad payers. Comparing to without the use of classifier to grant loans, the bank's revenue increased from \$40MM to \$217MM[8], contribute to a 443%[9] increase in revenue. The utility of the bank, good payers and bad payers are 213MM, 75.9MM and 14.9MM respectively. The total utility in this step is 304MM.

As times goes by, bad liars learn the methodology and characteristics that banks use to

[6]Booking rate = $\frac{774}{1000} = 77.4\%$

[7]Bank's revenue earned from good payers = 774K×\$1000×0.3 =232MM

[8]Bank's revenue = \$232MM - \$14.9MM = \$217.1MM

[9]Bank's increased in revenue = $\frac{217-40}{40} \times 100\% = 443\%$

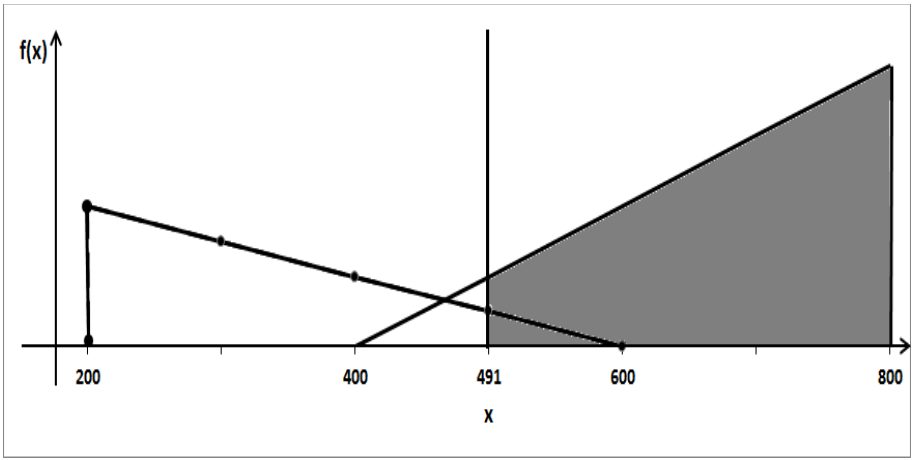


Figure 5.6: This plot presents the probability density function (PDF) of the scores of different borrowers. We set a to 200, A to 0 and η to 0 in Figure 5.3 and added the cut off line $\gamma = 491$. The dash line overlapping the solid line with circle patterns have scores ranging from 200 to 600 represents the PDF of bad honest borrowers and bad liars respectively. The solid line with scores ranging from 400 to 800 represents the PDF of good borrowers. The vertical line with $x = 491$ showed the cutoff value. The gray shaded region showed the applicants with their score greater than the cutoff value and will be granted loans.

grant loans. They will be motivated to add lies onto their attributes in order to increase their chances of loan approval. As discussed in Section 5.5.1, and under the assumptions in this chapter, the maximum amount of scores that bad liars can add onto their actual score is 200. In this scenario banks haven't put effort to check for lies and they keep using the same strategy to grant loans, therefore, η equals 0 and γ equals 491. To investigate the possibility of bad liars lie and the optimal amount of lies that they should add onto their attributes, we applied the parameters presented in Table 5.1, set $\eta = 0$ and $\gamma = 491$, varied A from 0 to 200 with stepsize 0.01 and for each values of A we calculate U_{Blied} as presented in Eq. (5.19). The plot of U_{Blied} versus A in Figure 5.7 showed that maximum utility of bad liars occurred when A equals to 0, which implies that using the parameters in this step, it is not profitable for bad liars to lie and they should set A to 0 in order to obtain their maximum utility of 7.44MM. Notice that the plot of U_{Blied} versus A in Figure 5.7 shows a concave up curve, stating that bad liars can only obtain maximum utility at the end points, either at $A = 0$ or at $A = 200$. Indeed, the optimal lie that bad liars should add onto their actual score follows Kolmogorov's zero-one law [40], stating that bad liars should either choose not to lie (set $A = 0$) or to produce maximum amount of lie (set $A = 200$).

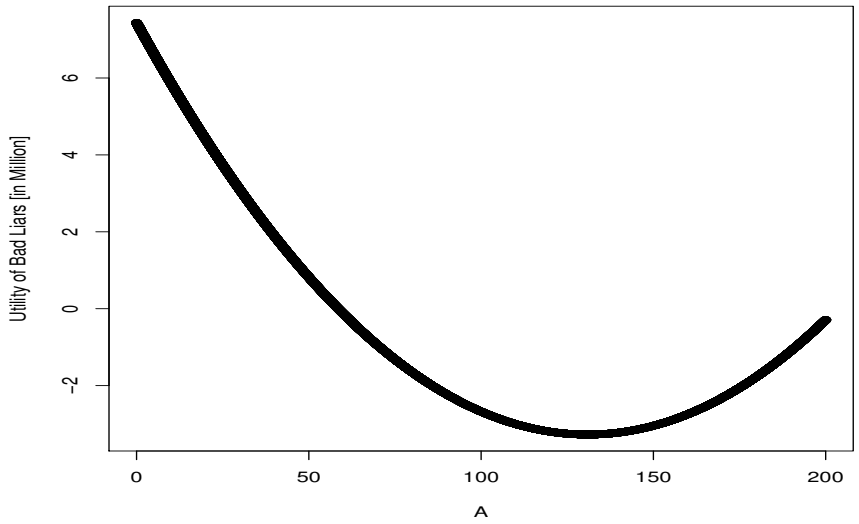


Figure 5.7: The above plot showed the utility of bad liars (in millions) versus different values of A. We applied the parameters in Table 5.1, set $\eta = 0$ and $\gamma = 491$, arbitrarily assigned h to \$3, varied A from 0 to 200 with stepsize 0.01, and for each values of A calculated U_{Blied}. This plot showed that maximum utility of bad liars occur at A equals 0, stating that bad liars should not any generate lie in order to maintain at their maximum utility level.

5.6.2 Scenario 2a: Economical to Lie But Not Economical to Check for Lies

This scenario represents the situation where it is economical for bad liars to lie but not for banks to check for lies. Figure 5.5 shows that this scenario has h less than \$2.61 and k can be anywhere between \$65.8 to \$80 [we did not compute results for $k >$ \$80]. In this case, bad liars can afford to add lies onto their actual attribute, however, it is not optimal for banks to spend money to check for lies since the unit cost to check for lies is too high. We arbitrarily set h to \$1 and k to \$70, evaluate all the objective functions as stated in Table 5.2 and presented the results in Table 5.4. Notice that the numbers in Step 1 of Table 5.3 and 5.4 are the same, it represents the situation where there are no liars in the lending business, and banks do not have to spend money to check for lies. In Step 2, bad liars learned the way to lie but still have to determine the appropriate amount of lies to add onto their attributes in order to escape from credit checks and obtain loans. Figure 5.8 is constructed much like Figure 5.7, only changing h from \$3 to \$1. The plot of U_{Blied} versus A in Figure 5.8 suggested that bad liars should set A to 200 in order to obtain their maximum utility of 39.7MM.

In addition, we substitute $a = 200$, $A = 200$, $\eta = 0$ and added the cutoff score $x = 491$ into Figure 5.3 to generate Figure 5.9. Notice that the triangles presented by the dash lines and the

	Step 1	Step2	Step3
η	0	0	0
A	0	200	200
γ	491	491	536
N_G	759K	759K	707K
N_B	7.44K	7.44K	2.53K
N_{Blied}	7.44K	59.7K	43.4K
N_T	774K	826K	753K
LiN_G	\$228MM	\$228MM	\$212MM
LN_B	\$-7.44MM	\$-7.44MM	\$-2.53MM
LN_{Blied}	\$-7.44MM	\$-59.7MM	\$-43.4MM
$Nk\eta$	\$0	\$0	\$0
U_{BK}	213MM	160MM	166MM
LIN_G	\$303MM	\$303MM	\$283MM
LiN_G	\$-228MM	\$-228MM	\$-212MM
U_G	75.9MM	75.9MM	70.7MM
U_B	7.44MM	7.44MM	2.53MM
LN_{Blied}	\$7.44MM	\$59.7MM	\$43.4MM
$(1-p_G)P_N NAh$	\$0	\$-20MM	\$-20MM
U_{Blied}	7.44MM	39.7MM	23.4MM
U_T	303MM	283MM	263MM

Table 5.4: Table of results for Scenario 2a where it is economical to lie but not to check for lies.

solid lines, representing bad true tellers and good payers respectively, have not changed when compared to Figure 5.6. The triangle presented by solid lines with circle patterns, representing bad liars, moved to the right and has scores ranging from 400 to 800. The increase in scores of bad liars is due to the added lies on their reported attributes. The shaded region in Figure 5.9 has a higher density compared to the shaded region in Figure 5.6, meaning that more applicants will be granted loans, consistent with the results presented in Table 5.4 that N_{Blied} increased by 702%[10], while N_T increased by 6.76%[11] from Step 1 to Step 2. The booking rate increased by 77.4% to 82.6%[12] from Step 1 to Step 2, contributed to an increase of roughly 5.22%. In normal situations, if there is no change in granting strategy or open credit, the number of loans granted in each period and booking rate should be about the same. An increase of 6.76% and 5.22% in the total number of loans granted and booking rate respectively from Step 1 to Step 2 should draw the bank's attention that some conditions might have changed and they may

[10]Number of loans granted to bad liars increased by $\frac{59.7-7.44}{7.44} \times 100\% = 702\%$

[11]Total loans granted increased by $\frac{826-774}{774} \times 100\% = 6.76\%$

[12]Booking rate = $\frac{826}{1000} \times 100\% = 82.6\%$

want to validate their existing granting strategy. Moreover, from the bank's point of view, the incurred losses from bad payers increased by 301%[13]. The overall utility of the bank decreased by 24.6%[14]. Although there is a cost of \$20MM for bad liars to make an appropriate lie, U_{Blied} in general increased by 434%[15]. The utility of good payers and bad honest payers remain the same as in Step 1, but the total utility decreased by 6.59%[16].

The primary remediation that the bank should consider is to update the classification model, which is under this framework, to recalibrate the cutoff score γ. We substitute $\eta = 0$, $A = 200$ and the parameters as presented in Table 5.1 into Eq. (5.10) to recalculate the cutoff score γ to be 536. We regenerated Figure 5.9 by moving γ from 491 to 536 and presented the new plot in Figure 5.10. Comparing Figure 5.9 and 5.10, the shaded region decreased in density, stating that the density of applicants with scores higher than the cutoff score has decreased. This result can also been seen in Step 3 of Table 5.4. The number of loans granted to good payers, bad honest payers, and bad liars have all been decreased. The total number of loans granted decreased by 8.82%[17]and booking rate decreased to 75.3%. The bank's revenue earned from good payers decreased by 7.30%[18] Step 1 to Step 3. The total incurred loss of the bank from bad payers decreased by 54%[19]. In general, the utility of the bank in Step 3 is higher than that in Step 2 but not as high as in Step 1. The utility of good payers, bad honest payers and bad liars all decreased. Note that the utility of bad liar is still relatively higher than that in Step 1. Overall utility in the lending environment decreased by 7.29%[20].

From the above results, it is obvious that changing the cutoff score of the classifier severely affects good and bad honest borrowers who have credit scores between 491 to 536. This group of borrowers have their loan request rejected due to the change of the cutoff score, and indeed is because of the existence of bad liars' lies. In addition, the utility of all parties decreases, banks should investigate other remediation process which can lower the utility of bad liars but maintain the utility of good payers and the bank to be at a similar level as stated in Step 1. We examine the possibility for banks to put effort to directly eliminate lies in applicants' attribute. We set $A = 200$, $k = \$70$ and varied η from 0 to 1 with step size 0.001. For each value of η, we first evaluate the corresponding γ by using Eq. (5.10), then calculate U_{BK} by applying the formula as stated in Eq. (5.16). Figure 5.11 presents the plot of U_{BK} versus η, which

[13]Bank's incurred loss from bad payers increased by $\frac{59.7+7.44-14.9}{14.9} \times 100\% = 301\%$
[14]Bank's utility decreased by $\frac{160.5-213}{213} \times 100\% = 24.6\%$
[15]Increased in bad liars' utility = $\frac{39.7-7.44}{7.44} = 434\%$
[16]Decreased in total utility = $\frac{283-303}{303} \times 100\% = 6.6\%$
[17]Decreased in total number of loans granted = $\frac{753-826}{826} \times 100\% = 8.82\%$
[18]Decreased in bank's revenue earned from good payers in Step 1 to Step 3 = $\frac{212-228}{228} \times 100\% = 7.30\%$
[19]Bank's incurred loss to bad payers = $\frac{(43.4+2.53)-(59.7+7.44)}{(59.7+7.44)} \times 100\% = -54\%$
[20]Decreased in total utility = $\frac{263-283}{283} \times 100\% = 7.29\%$

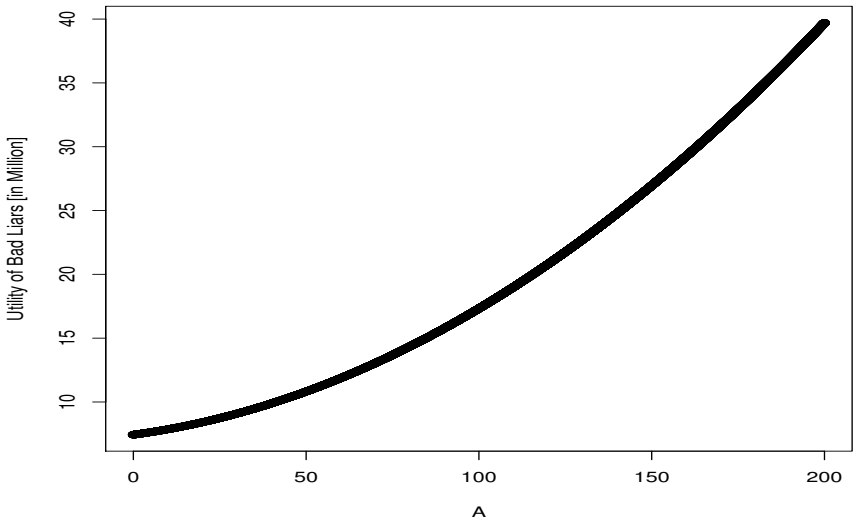


Figure 5.8: The above plot showed the utility of bad liars (in millions) versus different values of A. We applied the parameters in Table 5.1, set $\eta = 0$ and $\gamma = 491$, arbitrarily assigned h to \$1, varied A from 0 to 200 with step size 0.01, and for each values of A calculated U_{Blied}. This plot showed that maximum utility of bad liars occur at A equals 200 , stating that bad liars should lie at their maximum to obtain at their maximum utility level.

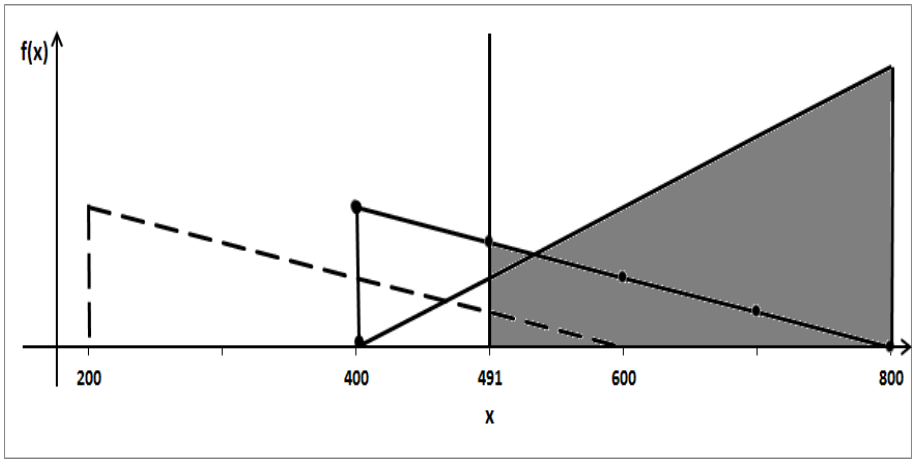


Figure 5.9: This plot presents the probability density function (PDF) of the attributes of different borrowers. We set a to 200, A to 200 and η to 0 in Figure 5.3 and added the cut off line $\gamma = 491$. The dash line with attribute values ranging from 200 to 600 represents the PDF of bad honest borrowers. The solid line with circle patterns with attribute values ranging from 400 to 800 represents the PDF of bad liars, while the solid line with attribute values ranging from 400 to 800 represents the PDF of good borrowers. The gray shaded region showed the applicants with their attribute value greater than the cutoff value and will be granted loans.

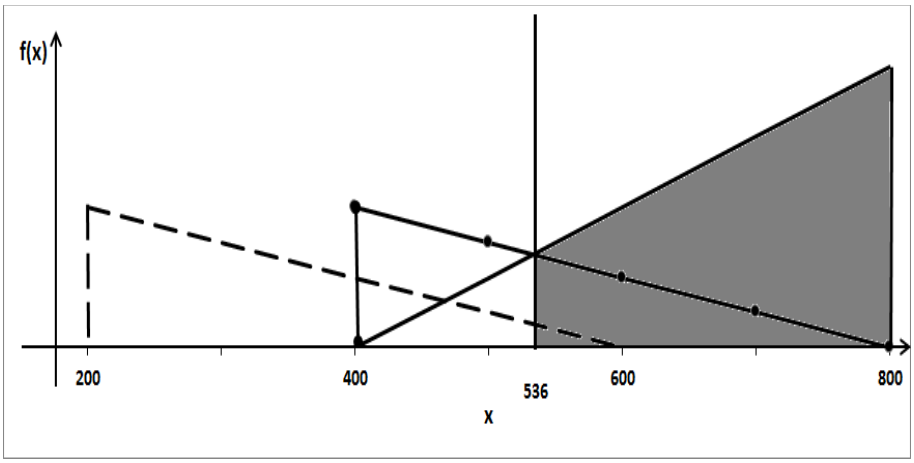


Figure 5.10: The cutoff value has been recalculated by substituting $A = 200$, $\eta = 0$ into Eq. (5.10) and resulted with $\gamma = 536$. This plot presents the probability density function (PDF) of the attributes of different borrowers. We set a to 200, A to 200 and η to 0 in Figure 5.3 and added the cut off line $\gamma = 536$. The dash line with attribute values ranging from 200 to 600 represents the PDF of bad honest borrowers. The solid line with circle patterns with attribute values ranging from 400 to 800 represents the PDF of bad liars, while the solid line with attribute values ranging from 400 to 800 represents the attributes of good borrowers. The gray shaded region showed the applicants with their attribute value greater than the cutoff value and will be granted loans.

shows a monotone decreasing curve, stating that the bank should set η to 0 in order to maintain at its maximum utility of 166MM. In this scenario, banks should not put effort to check for lies in applicants' attribute due to the fact that the unit cost to check for lies is too high. In the succeeding scenarios, we will lower the unit cost to check for lies, examine the effect of directly checking liars, and determine the optimal amount of effort that banks should spend to check for lies.

5.6.3 Scenario 2b: Always Economical to Lie and Check for Lies

This scenario represents the situation where it is always economical for bad liars to lie and for banks to check for lies. The white region in Figure 5.5 shows the values of h and k which contribute to this scenario. To illustrate this scenario, we keep $h = \$1$ which is the same as in Scenario 2a, arbitrarily set $k = \$50$ and presented the results in Table 5.5. Notice that all the values in Steps 1, 2, and 3 stated in Table 5.5 are the same as in Table 5.4. The first three steps in both the scenarios have the same results. Again, Step 1 shows the original lending environment where there are no liars in the industry, and banks do not have to spend extra effort to verify borrowers' attributes. In Step 2, since the unit cost to lie (h) was set to be \$1, liars can afford to lie, it presents the environment where liars started to add lies onto their attributes, successfully

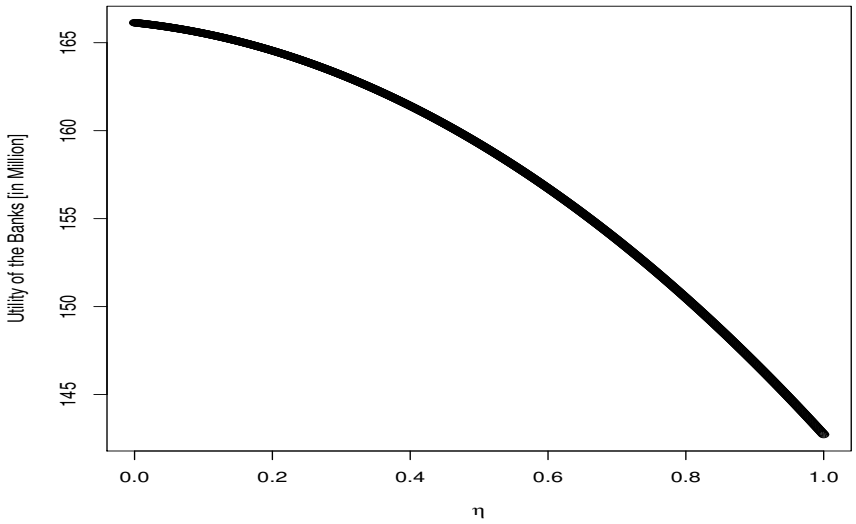


Figure 5.11: The above plot showed the utility of the bank (in millions) versus different values of η. We applied the parameters in Table 5.1, set $A = 200$, assigned k to \$70, varied η from 0 to 1 with step size 0.001, and for each values of η calculated U_{BK}. This plot showed that maximum utility of the bank occur at η equals 0 , stating that banks should not spend money to check for lies in applicants' attribute.

escaped from regular credit checks and obtained loans. The shift in the distribution of bad liars' credit score lead to increase in bank's losses. Banks will then update their classification model so as to equalize risk in Step 3. However, when comparing bank's utility in Step 1, changing the classifier alone does not generate a high enough utility. Therefore, in Step 4, banks consider to directly put effort to check for lies in applicants' attribute, indeed, they determine the optimal amount of effort that should be included on top of their regular credit check process which can minimize cost and maximize return.

In Step 4, given the initial conditions in Table 5.1, banks want to determine whether it is profitable to put extra effort to check for lies in applicants' attribute. We apply the parameters presented in Table 5.1, set $A = 200$, $k = \$50$ and varied η from 0 to 1 with step size 0.001. For each value of η we calculate the corresponding U_{BK}. Figure 5.12 presents the plot of U_{BK} versus η and suggests banks to set η to 0.412 in order to obtain the maximum utility of 170MM. Although it costs the bank \$20.6MM over and above their regular credit check process, it lowers the bank's incurred losses to bad payers by 56.5%[21] and increases bank's utility by 1.97%[22]. We then substitute $\eta = 0.412$ and $A = 200$ into Eq. (5.10) to update the

[21] With extra effort added to check for lies in borrowers' reported attributes, bank's incurred losses lower by $\frac{(4.24+25.0)-(7.44+59.7)}{(7.44+59.7)} \times 100\% = 56.5\%$

[22] Banks' increased in utility when 0.412 afford was added to check for lies: $\frac{169-166}{166} \times 100\% = 1.97\%$

	Step1	Step2	Step3	Step4
η	0	0	0	0.412
A	0	200	200	200
γ	491	491	536	518
N_G	759K	759K	707K	731K
N_B	7.44K	7.44K	2.53K	4.24K
N_{Blied}	7.44K	59.7K	43.4K	25.0K
N_T	774K	826K	753K	760K
LiN_G	\$228MM	\$228MM	\$212MM	\$219MM
LN_B	\$-7.44MM	\$-7.44MM	\$-2.53MM	\$-4.24MM
LN_{Blied}	\$-7.44MM	\$-59.7MM	\$-43.4MM	\$-25.0MM
$Nk\eta$	\$0	\$0	\$0	\$-20.60MM
U_{BK}	213MM	160MM	166MM	169MM
LIN_G	\$303MM	\$303MM	\$283MM	\$292MM
LiN_G	\$-228MM	\$-228MM	\$-212MM	\$-219MM
U_G	75.9MM	75.9MM	70.7MM	73.1MM
U_B	7.44MM	7.44MM	2.53MM	4.24MM
LN_{Blied}	\$7.44MM	\$59.7MM	\$43.4MM	\$25.0MM
$(1-p_G)P_NNAh$	\$0	\$-20MM	\$-20MM	\$-20MM
U_{Blied}	7.44MM	39.7MM	23.4MM	4.99MM
U_T	303MM	283MM	263MM	252MM

Table 5.5: Table of results for Scenario 2b where it is always economical to lie and check for lies.

cutoff score of the classifier to 518. Notice that putting effort $\eta = 0.412$ to check for lies will change the classifier to generate a smaller cutoff score when compared to Step 3. Figure 5.13 showed the PDF of different borrowers and the cutoff score in Step 4. According to Figure 5.4, the credit scores of bad liars vary from $a + A(1-\eta)$ to $3a + A(1-\eta)$, substituting $A = 200$ and $\eta = 0.412$ resulted in having bad liars' scores vary from 318 to 718 in Step 4. The additional effort η shifted the PDF of bad liars and the cutoff score of the classifier more towards the left. Notice that the addition of lies and extra effort to check for lies in the lending business did not change the PDF of good borrowers and bad honest borrowers. However, shifting the cutoff line more towards the left increased the number of loans granted to good and bad honest borrowers. As one can see in Table 5.5, the number of loans granted in Step 4 is higher than that in Step 3, except for bad liars that the number of loans granted decreased by 42.5%[23] and actually the number of loans granted to bad liars reduced by 58.2%[24] when compared to Step 2. Therefore, the effect of spending η amount of effort to directly check for lies reduces 58.2% of

[23]Number of loans granted to bad liars decreased by $\frac{25.0-43.4}{43.4} \times 100\% = 42.5\%$.

[24]Number of granted to bad liars reduced by $\frac{25.0-59.7}{59.7} \times 100\% = 58.2\%$

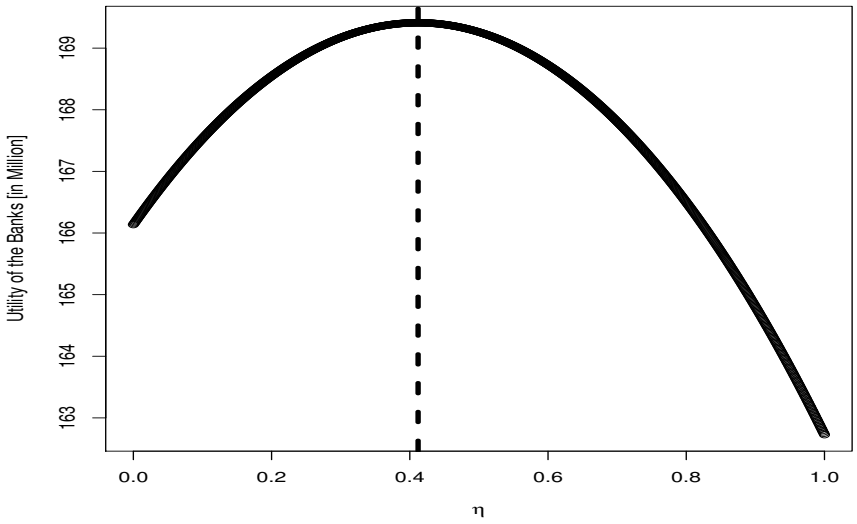


Figure 5.12: The above plot showed the utility of the banks (in millions) versus different values of η. We applied the parameters in Table 5.1, set $A = 200$, $k = \$50$, and varied η from 0 to 1 with step size 0.001. For each values of η we calculated the corresponding U_{BK}. The dash line showed that maximum utility of the banks occurs at $\eta = 0.412$ stating that banks should put 0.412 amount of effort to check for lies in applicants' attribute.

the loans granted to bad liars. Adding η amount of effort to check for lies in applicants' attribute requires a cost of \$20.6MM. However, banks still obtained at a utility level of 170MM which is the highest amongst Step 2 to Step 4. Since there are more good and bad honest borrowers granted loans, the utility of good and bad honest borrowers are also higher when compared to Step 3. The introduction of η provides a mixed effect to bad liars utility. While it shifts the PDF more towards the left which decreases the number of loans granted to bad liars, it also moves the cutoff score more towards the left, which indeed increases the total number of loans granted. Overall, the utility of bad liars decreased by 78.7% [25] with the introduction of η. The mixed effect of η and A makes the lending environment not as efficient and contributed 17.1%[26] decrease in overall utility when compared to Step 1.

With the extra η amount of effort added onto banks' regular credit check process, bad liars have to revise their strategy to obtain loans. In particular, they have to reconsider whether it is still profitable for them to lie and may have to alter the amount of lie that should be added onto their actual attribute. We apply the parameters in Table 5.1, set $\eta = 0.412$, $\gamma = 518$, $h = \$1$ and varied A from 0 to 200 with step size 0.001, and for each value of A, we calculate the corresponding U_{Blied} using Eq. (5.19). Figure 5.14 shows the plot of U_{Blied} versus A. This plot

[25] Utility of bad liars decreased by $\frac{4.99-23.4}{23.4} \times 100\% = 78.7\%$

[26] Overall utility decreases by $\frac{252-303}{303} \times 100\% = 17.1\%$.

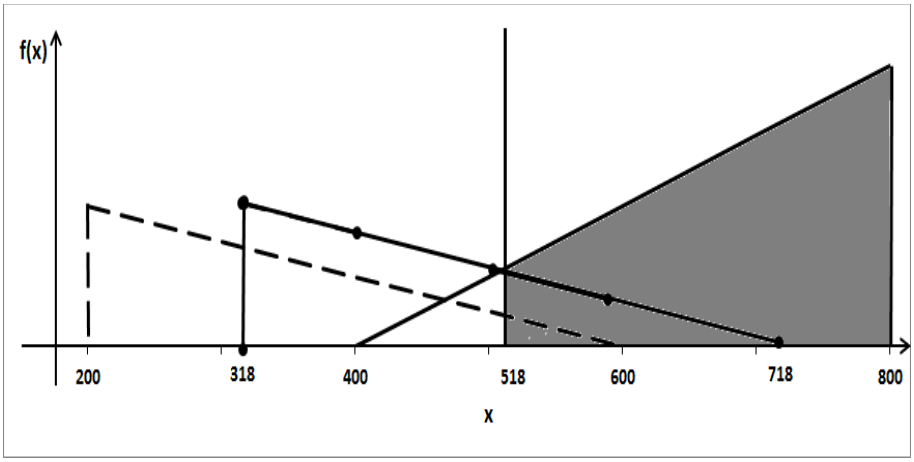


Figure 5.13: This plot shows the PDF of different borrowers in Step 4. The cutoff value has been calculated by substituting $A = 1$, $\eta = 0.412$ into Eq. (5.10) and resulted with $\gamma = 2.59$. This plot presents the probability density function (PDF) of the scores of different borrowers. We set a to 1, A to 1 and η to 0.412 in Figure 5.4 and added the cut off line $\gamma = 2.59$. The dash line with scores ranging from 1 to 3 represents the PDF of bad honest borrowers. The solid line with circle patterns with scores ranging from 1.59 to 3.59 represents the PDF of bad liars, while the solid line with scores ranging from 2 to 4 represents the scores of good borrowers. The gray shaded region showed the applicants with their scores greater than the cutoff value and will be granted loans.

suggests that bad liars should set A to 200 in order to maintain at their maximum utility level of 4.99MM. Therefore, bad liars should continue to lie, while banks should continue to check for lies in this scenario.

5.6.4 Scenario 2c: Economical to Lie But Check for Lies Periodically

This scenario represents the situation where it is economical for bad liars to lie but banks should only periodically check applicants' attributes for lies. The black shaded region in Figure 5.5 shows the values of h and k which contribute to this scenario. To illustrate this scenario, we keep $h = \$1$ same as in Scenario 2a and 2b, arbitrarily set $k = \$20$ and presented the results of each step in Table 5.6. Notice that setting h to 1 contributes to having all the values in Steps 1, 2 and 3 to be the same as in Table 5.4 and 5.5. The analysis of those three steps are exactly the same as in previous scenarios and will not be repeated here. Note that this scenario assumes the unit cost required to check for lies are lower than in other scenarios, which indeed increases the affordability of the banks. In Step 4, we applied the parameters in Table 5.1, set $A = 200$, $k = \$20$, and varied η from 0 to 1 with step size 0.001. For each values of η we calculated the corresponding U_{BK}. The plot of U_{BK} versus η in Figure 5.15 suggests banks to set η to 1 and

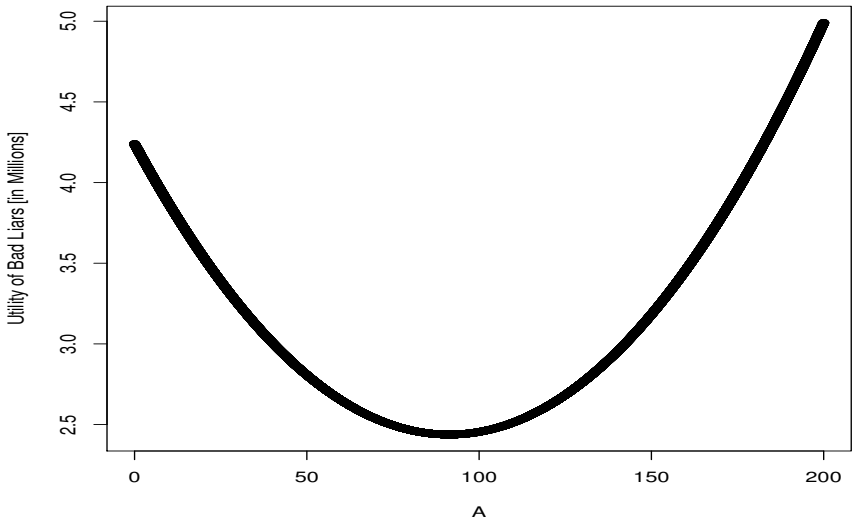


Figure 5.14: This plot showed the utility of bad liars (in millions) versus different values of A in Step 4 of Scenario 2b. We applied the parameters in Table 5.1, set $\eta = 0.412$ and $\gamma = 518$, arbitrarily assigned h to \$1 and k to \$50, and varied A from 0 to 200 with step size 0.01. For each values of A we calculated U_{Blied}. This plot showed that maximum utility of bad liars occurred at A equals 1 , stating that bad liars should keep adding maximum amount of lies onto their actual score in order to obtain at a utility level of 4.99MM.

put full amount of effort to check for lies in applicants' attribute in order to obtain 193MM of utility.

In this scenario, bad liars lied at an maximum amount and banks put full effort to eliminate all the lies in applicants' reported attributes, we substitute both η to 1 and A to 200 into Eq. (5.10) to evaluate the cutoff score of the classifier to be 491 as presented in Step 4 of Table 5.6. Notice that when both η and A are set at their maximum value, the cutoff score and the distribution of the credit scores of bad liars are the same value as in Step 1. Indeed, the probability density functions of all the borrowers and the cutoff score in Step 4 is the same as in Step 1 and is presented in Figure 5.6. The Bank's strategy of putting 100% of effort to validate applicants' attribute can fully eliminate all the lies and correct bad liars reported attribute to its actual value. In reality, it is very difficult to clean data and the chances of having data that does not contain any noise is very low, however, it is not impossible to have all the lies removed in the dataset and our research captures this.

In Step 4, banks put full amount of effort to remove all the lies in the data. The distribution of borrowers' attribute and the cutoff score of the classifier remains the same as in Step 1. Therefore, the number of loans granted to borrowers remains the same as in Step 1. However, it costs the banks an addition of \$20MM in their credit check process but contribute to an

	Step1	Step2	Step3	Step4	Step5	Step6
η	0	0	0	1	1	0
A	0	200	200	200	0	0
γ	491	491	536	491	491	491
N_G	759K	759K	707K	759K	759K	759K
N_B	7.44K	7.44K	2.53K	7.44K	7.44K	7.44K
N_{Blied}	7.44K	59.7K	43.4K	7.44K	7.44K	7.44K
N_T	774K	826K	753K	774K	774K	774K
LiN_G	\$228MM	\$228MM	\$212MM	\$228MM	\$228MM	\$228MM
LN_B	\$-7.44MM	\$-7.44MM	\$-2.53MM	\$-7.44MM	\$-7.44MM	\$-7.44MM
LN_{Blied}	\$-7.44MM	\$-59.7MM	\$-43.4MM	\$-7.44MM	\$-7.44MM	\$-7.44MM
$Nk\eta$	\$0	\$0	\$0	\$-20MM	\$-20MM	\$0
U_{BK}	213MM	160MM	166MM	193MM	193MM	213MM
LIN_G	\$303MM	\$303MM	\$283MM	\$303MM	\$303MM	\$303MM
LiN_G	\$-228MM	\$-228MM	\$-212MM	\$-228MM	\$-228MM	\$-228MM
U_G	75.9MM	75.9MM	70.7MM	75.9MM	75.9MM	75.9MM
U_B	7.44MM	7.44MM	2.53MM	7.44MM	7.44MM	7.44MM
LN_{Blied}	\$7.44MM	\$59.7MM	\$43.4MM	\$7.44MM	\$7.44MM	\$7.44MM
$(1-p_G)P_NNAh$	\$0	\$-20MM	\$-20MM	\$-20MM	\$0	\$0
U_{Blied}	7.44MM	39.7MM	23.4MM	-12.6MM	7.44MM	7.44MM
U_T	303MM	283MM	263MM	263MM	283MM	303MM

Table 5.6: Table of results for Scenario 2c where it is economical to lie and check for lies periodically.

increased of 16%[27] in utility. The utility of good borrowers and bad honest borrowers remain the same as in Step 1. It costs \$20MM for bad liars to produce lies, the success on removing all the lies in the data canceled the effect of lying, and this contributes to generate negative utility for bad liars. The total utility in the lending business in Step 4 is slightly higher than in Step 3 but still lower than that in Step 1.

The negative utility that bad liars generate in Step 4 provided them a sign that it is necessary to reconsider whether it is beneficial for them to lie. We regenerate the plot of U_{Blied} versus A by reassigning η to 1 and γ to 491 as calculated in Step 4, varied A from 0 to 200 with step size 0.01, and for each values of A substitute the fixed parameters in Table 5.1 into Eq.(5.19) to calculate U_{Blied}. Figure 5.16 showed that U_{Blied} decreases monotonically with A, suggesting that bad liars should reset A to zero and be honest on reporting their attributes. Therefore, we set $A = 0$, re-evaluated all the objective functions and presented the results in Step 5 of Table 5.6. Note that this step shows the situation where bad liars did not lie but banks still

[27] Bank's increased in utility = $\frac{192-166}{166} \times 100\% = 16\%$

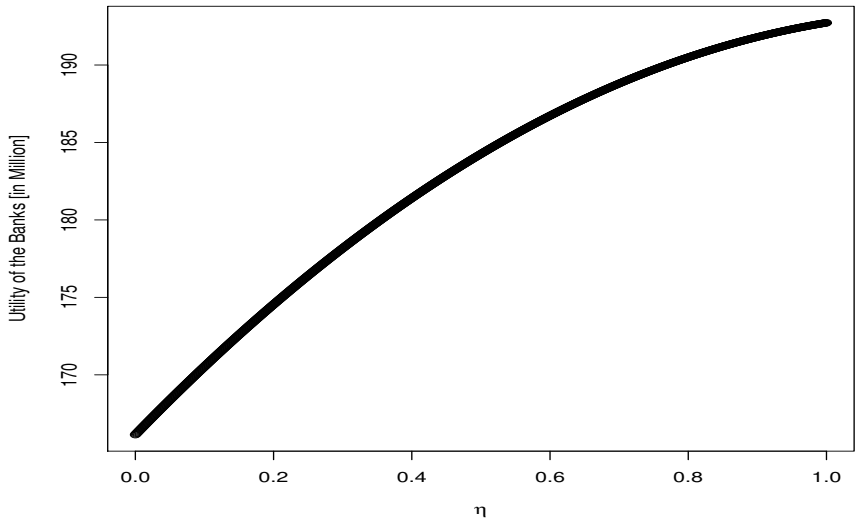


Figure 5.15: The above plot showed the utility of the banks (in millions) versus different values of η. We applied the parameters in Table 5.1, set $A = 200$, $k = \$20$, and varied η from 0 to 1 with step size 0.001. For each values of η we calculated the corresponding U_{BK}. This plot suggested banks to set η to 1 to put full effort to check for lies in applicants' attribute. The maximum amount of utility that banks can obtained is 193MM.

put full effort to check for the accuracy of applicants' attribute. All the objective functions in Step 5 have the same value as in Step 4 except for U_{Blied} and U_T. This step stated that banks can consider to permanently put full effort to check for lies in applicants' attribute as a precaution. Then banks' utility will always maintain at 193MM and will never experience bad liars' shocked, in which case will lower the utility to 161MM as stated in Step 2. In addition, if banks can further investigate different methods to detect specifically on the time that bad liars added lies onto their attribute, they can put effort to check for lies only at those time that bad liars did lied. Banks can then always generate maximum utility. As in Step 5, bad liars did not add lies onto their attribute, banks can then minimize cost and remove the effort spend on checking lies, this will then make all the objective functions to have the same value as in Step 1. We specifically presented the numbers again in Step 6 to show the entire cycle in this scenario.

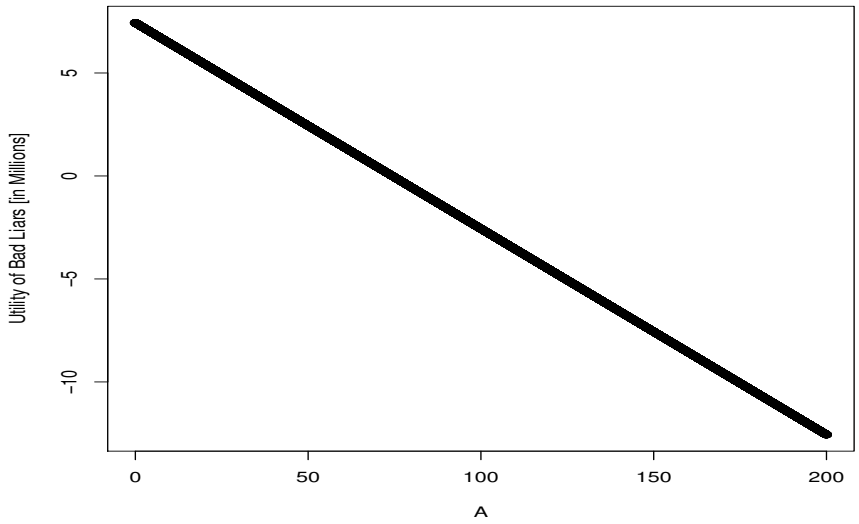


Figure 5.16: This plot showed the utility of bad liars (in millions) versus different values of A in Step 5 of Scenario 2c. We applied the parameters in Table 5.1, set $\eta = 1$ and $\gamma = 491$, assigned h to \$1 and k to \$20, and varied A from 0 to 200 with step size 0.01. For each values of A we calculated U_{Blied}. This plot showed that maximum utility of bad liars occurred at A equals 0 , stating that bad liars should not carry on to lie, instead, they will be better of to report their actual attribute.

5.7 Conclusions

In this chapter, we extended the idea presented in Chapter 4 to model a game theoretic setting which banks and bad liars select their parameters in competition with each other. In particular, we examined the strategic impact of the unit cost for banks to check for lies and the unit cost to bad liars involved in adding lies to their attributes. We showed that if the unit cost of lying is too high, bad liars will not be able to afford to lie and banks will obtain maximum utility. In cases where the unit cost to lie is affordable, bad liars can choose a lie intensity which maximize their utility, while banks can attempt to modify the classifier to avoid granting loans to bad liars. However, this also lowered the number of loans granted to good borrowers, which in turn decreased the utility of good borrowers and affects the profitability of the banks. A better strategy for the banks is to directly exert effort to eliminate lies. Banks will then determine the optimal amount of effort to spend on eliminating lies, which affects their cost and profit. The amount of effort that banks can afford to check for lies depends on bad liars' lie intensity and the cost require to eliminate lies. In cases where it is too expensive to eliminate lies, banks must live with bad liars. While in cases where banks can afford to eliminate lies, we showed that it is possible for bad liars to continue to lie, while banks continue to eliminate lies, both

parties will exhibit sub-optimal conditions. On the other hand, it is also possible that bad liars add lies and banks eliminate lies in a periodic cycle. The time period over which bad liars should lie and banks need to eliminate lies will be of interest in our further studies.

We plan to link our work to other related economic literature. First, the problem of optimal auditing has been studied by Dionne *et al.* [41], who found the optimal time to audit an insurance claim when some fraud indicators are observed. This could be applied to determine the optimal time period to check for lies.

Second, dynamic adverse selection modeling methods as studied by Chatterjee and Ionescu [42], where the impact of college students borrowed money to attend school but did not manage to graduate, can be applied to study the impact of bad borrowers who lied, but were not able to obtain loans.

In addition, it must be noted that in the model presented here, liars 'acted first' in that they chose their effort level to which the banks then responded. In future work, it would be interesting to see what happened if a different model, in which banks moved first, was considered.

Bibliography

[1] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens and J. Vanthienen, Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, **54**(6), (2003), 627-635.

[2] B. Liu, E. Roca, What drives mortgage fees in Australia?, *Accounting & Finance*, (2014). DOI: 10.1111/acfi.12068.

[3] E.I. Altman, Financial Ratios, Discriminant analysis and the prediction of corporate bankruptcy, *The Journal of Finance*, **23**(4) (1968), 589-609.

[4] K. Dejaeger, B. Hamers, J. Poelmans and B. Baesens, A novel approach to the evaluation and improvement of data quality in the financial sector, *Proceedings of the International Conference on Information Quality (ICIQ 2010)*, Little Rock, AR, United States, 2010.

[5] Lessmanna, Stefan, H. Seow, B. Baesens, and L. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update, *In Credit Research Centre*, Conference Archive. 2013.

[6] L. Thomas, A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers, *International Journal of Forecasting*, **16**(2), (2002), 149-172.

[7] L. Thomas, D. B. Edelman and J. N. Crook, *Credit Scoring and Its Applications*, SIAM, Philadelphia, USA, 2002.

[8] R. Anderson, *The Credit Scoring Toolkit: Theory and practice for retail credit risk management and decision automation*, Oxford University Press, Oxford, UK, 2007.

[9] A. Frankel, Prime or Not So Prime? An Exploration of US Housing Finance in the New Century, *BIS Quarterly Review*, pp. 67-78.

[10] Y. Kim, S.B. Kang and J.I. Seo, Bayesian Estimations on the Exponentiated Half Triangle Distribution Under Type-I Hybrid Censoring, *Journal of the Korean data & Information Science Society*, **22**(3), (2011), 565-574.

[11] Y.W. Chien and S.A. Devaney, The Effects of Credit Attitude and Socioeconomic Factors on Credit Card and Installment Debt, *Journal of Consumer Affairs*, **35**(1), (2001), 162-179, DOI: 10.1111/j.1745-6606.2001.tb00107.x.

[12] T. Wang, Paying Back to Borrow More: Reputation and Bank Credit Access in Early America, *Explorations in Economic History*, **45**(4), (2008), 477-488.

[13] J. Woo, Reliability In a Half-Triangle Distribution and a Skew-Symmetric Distribution, *Journal of Korean Data & Information Science Society*, **18**(2), (2007), 543-552.

[14] M. Chong, C. Bravo and M. Davison, How Much Effort Should be Spent to Detect Fraudulent Applications When Engaged in Classifier-Based Lending?, *Intelligent Data Analysis*, **19**(S1), (2015), S87-S101.

[15] T. Simpson, A Letter to the Right Honorable George Earl of Macclesfield, President of the Royal Society, on the Advantage of Taking the Mean of a Number of Observations, in Practical Astronomy, *Philosophical Transactions of the Royal Society of London*, **49**, (1755), 8293.

[16] R. Schmidt, Statistical analysis of one-dimensional distributions, *Annals of Mathematical Statistics*, **5**, (1934), 3043.

[17] A.S.K. Ayyangar, The Triangular Distribution, *Mathematics Student*, **9**, (1941), 85-87.

[18] N.L. Johnson and S. Kotz, Non-Smooth Sailing or Triangular Distributions Revisited After Some 50 Years, *Journal of the Royal statistical Society*, **48**(2), (1999), 179-187.

[19] D. Johnson, The triangular distribution as a proxy for the beta distribution in risk analysis, *Journal of the Royal Statistical Society: Series D (The Statistician)*, **46**(3), (1997), 387-398.

[20] D. Johnson, Triangular approximations for continuous random variables in risk analysis, *Journal of the Operational Research Society*, (2002), 457-467.

[21] C.E. Clark, Letter to the Editor-The PERT Model for the Distribution of an Activity Time, *Operations Research*, **10**(3), (1962), 405-406.

[22] F.E. Grubbs, Letter to the Editor-Attempts to Validate Certain PERT Statistics or "Picking on PERT, *Operations Research*, **10**(6), (1962), 912-915.

[23] K.R. MacCrimmon and C.A.Ryavec, An analytical study of the PERT assumptions, *Operations Research*, **12**(1), (1964), 16-37.

[24] J.J. Moder and E.G. Rodgers, Judgment estimates of the moments of PERT type distributions, *Management Science*, **15**(2), (1968), B-76.

[25] I. Văduva, On Solving some types of Multiple Attribute Decision Making Problems, *Romanian Journal of Economic Forecasting*, **15**(1), (2012), 41-61.

[26] T.M. Williams, Practical use of distributions in network analysis, *Journal of the Operational Research Society*, (1992), 265-270.

[27] D.L. Keefer and S.E. Bodily, Three-point approximations for continuous random variables, *Management Science*, **29**(5), (1983), 595-609.

[28] M.J.Courchane, B.J. Surette and P.M. Zorn, Subprime Borrowers: Mortgage Transitions and Outcomes, *The Journal of Real Estate Finance and Economics*, **29**(4), (2004), 365-392.

[29] S. Antle, A Practical Guide to Section 347 of the Criminal Code? Criminal Rates of Interest, *Can. Bus. LJ*, **23**, 323.

[30] Basel Committee on Banking Supervision, December 2010. Basel III: A global regulatory framework for more resilient banks and banking systems. Tech. rep., Bank for International Settlements.

[31] J.A. Roberts and E. Jones, Money attitudes, credit card use, and compulsive buying among American college students, *Journal of Consumer Affairs*, **35**(2), (2001), 213-240.

[32] M. Csikszentmihalyi and E. Rochberg-Halton, *The Meaning of Things*, New York: Cambridge University Press, 1981.

[33] Equifax Canada Inc., `http://www.newswire.ca/news-releases/equifax-launches-citadeltm-the-most-advanced-application-fraud-prevention-and-r html`, (2010-10).

[34] M. Kim and P. Vieira, Canada Banking Regulator Tightens Mortgage-Lending Oversight, *The Wall Street Journal*, `http://www.wsj.com/articles/canada-banking-regulator-tightens-mortgage-lending-oversight-1467907512`, (2016-07)

[35] M. Malekipirbazari and V. Aksakalli, Risk assessment in social lending via random forests, *Expert Systems with Applications*, **42**(10), (2015), 4621-4631.

[36] S. J. Bailey, D. Asenova and J. Hood, Making widespread use of municipal bonds in Scotland?, *Public Money & Management*, **29**(1), (2009), 11-18.

[37] R. A. Easterlin, Does money buy happiness?. *The Public Interest*, (30), (1973), 3.

[38] L. C. Fernald, R. Hamad, D. Karlan, E. J. Ozer, and J. Zinman, Small individual loans and mental health: a randomized controlled trial among South African adults, *BMC public health*,**8**(1), (2008), 1.

[39] J. J. Walczyk, J. P. Schwartz, R. Clifton, B. Adams, M. I. N, Wei and P. Zha, Lying person-to-person about life events: a cognitive framework for lie detection, *Personnel Psychology*, **58**(1), (2005), 141-170.

[40] Wikipedia, Kolmogorov's zero-one law, `https://en.wikipedia.org/wiki/Kolmogorov's_zero?one_law`.

[41] G. Dionne, F. Giuliano and P. Picard, Optimal auditing with scoring: Theory and application to insurance fraud, *Management Science*, **55**(1), (2009), 58-70.

[42] S. Chatterjee and F. Ionescu, Insuring student loans against the financial risk of failing to complete college, *Quantitative Economics*, **3**(3), (2012), 393-420.

Chapter 6

Conclusions

6.1 Summary

Nowadays everyone needs credit. The applications of credit scoring techniques can benefit financial institutions and borrowers alike. Financial institutions employ credit scoring techniques to make lending decisions, which contribute to higher efficiency, lower cost, minimum risk, and maximum profit. Borrowers also benefit from quicker access to credit. They learn about the methods and skills that analysts use to offer credits, which can increase the chance of getting loans. They can acquire a better understanding about how their characteristics will be gathered, analyzed and transformed into a credit score which can reflect their credit worthiness. They can also seek ways to behave in such a way so as to maintain a good credit history.

Credit fraud, improperly handled, can cause inaccurate loan decisions which contribute to suboptimal loan book profitability. In this thesis, we study the effect on consumer lending if some borrowers strategically falsify one or more of their attributes on their application form. We applied discriminant analysis, a well-known classification method that has been widely used in credit scoring, to make granting decisions. Below is a brief summary of what we have accomplished in the chapters containing the research dealt with in this thesis:

Chapter 2 (which is a review of existing techniques) explored different methods use in the context of credit scoring. We specifically examined the background of discriminant analysis, and analyzed the prediction power of this method by assuming the characteristics of borrowers follows the normal distribution. In addition, we provided an overview of logistic regression analysis, another well-known method used in credit scoring. We concluded that discriminant analysis, a model-driven approach, is more suitable for the theoretical exploration of models in our research and will be used as the modeling method throughout this thesis.

In Chapter 3 (also published as [1]), we studied the effect of credit fraud on model accuracy. Analysts believe that using more attributes to develop credit scoring models will always

increase the model predictive power. We showed that if some borrowers fraudulently falsified their attributes on their application form, using higher dimensional data to develop credit scoring models may result in a more complex model with lower predictive power when compared to a model built using a dataset with lower dimensions. This draws our interest in the study of credit fraud in later chapters.

Chapter 4 (or the related paper [2]) presented a simple stylized model to investigate how banks respond to bad borrowers' lies. From the bank's perspective our study shows that, while there are cases in which huge lies are better left alone, intelligent effort spent on reducing lies usually increases revenue. The increased amount of profit depends on the cost spent to reduce lies and the lie intensity which bad borrowers added to their reported attributes. The example discussed in the chapter suggests that, if eliminating lies is not expensive, banks should reduce the impact of lies only until they are small enough that the classifier can take over and itself distinguish good from bad borrowers. From the borrowers' point of view, we showed that it is possible for bad borrowers to lie cleverly to pass credit checks and obtain loans. In fact, liars have a strong incentive to behave strategically, if bad liars know the attribute and the type of model used to classify prospective borrowers, we showed that it is possible for clever bad liars to select an optimal lie intensity which minimize their cost of lying but maximize their chance of obtaining loans.

In Chapter 5, we extended the idea presented in Chapter 4 to model a game theoretic setting which banks and bad liars select their parameters in competition with each other. In particular, we examined the strategy impact of the unit cost for banks to check for lies and the unit cost requires bad liars to add lies onto their attributes towards their choices on selecting their parameters. We showed that if the unit cost of lying is too high, bad liars will not be able to afford to lie and banks will obtain maximum utility. In cases where the unit cost to lie is affordable, bad liars can choose a lie intensity which maximize their utility, while banks can determine the amount of effort to spend on eliminating lies, which indeed affects their cost and profit. The amount of effort that banks can afford to check for lies, depends on bad liars' lie intensity and the cost require to eliminate lies. If the cost to eliminate lies is too high, banks can only live with bad liars. While in cases where banks can afford to eliminate lies, we showed that it is possible for bad liars to continue to lie, while banks continue to eliminate lies, both parties will exhibit sub-optimal conditions. On the other hand, it is also possible that bad liars add lies and banks eliminate lies in a periodic cycle. The time period over which bad liars should lie and banks need to eliminate lies will be of interest in our further studies.

6.2 Business Insights from this work

Credit scoring is used to provide a quick, effective, and unbiased way of granting loans to applicants. Unfortunately, the questions used to elicit the numerical values of attributes can be gamed by applicants who wish to obtain access to loans for which they would not qualify if they supplied all answers truthfully. This thesis discusses some of the aspects of this game between borrower and lender.

From the lenders point of view, efforts might be devoted to correcting the numerical value of attributes, without penalizing borrowers for supplying incorrect attribute values. That is the approach taken here. In this case there is a tradeoff between the improved decision enabled by correcting attributes and the cost of correcting. One business insight here is that attributes that are easy and cheap to verify might be a better basis for scoring than more information rich, yet more costly to verify, attributes. Thus, for instance, the amount of salary earned by a waged employee might be a better attribute than the value of collateral, which requires subjective appraisal. This thesis discusses from the lenders point of view, the optimal effort to expend on checking and correcting attributes. This in turn requires an estimate of the size of lies used (the "A" parameter.) In addition, the cost required to check for lies is also a fact that lenders need to consider.

However, now the game aspect arises. If the borrower knows the attribute being checked and the model the lender is using for the size of lies added to the attribute by dishonest borrowers and the cost of correcting the attribute, this thesis shows that the borrower can select a different lie amount to minimize their cost of lying and maximize their profit. The conclusion from this is that lenders should keep their methodology secret. This might mean asking questions about attributes not being used in the credit model, and certainly keeping scoring algorithms confidential. However, over time, this kind of information is bound to leak out. That suggests changing scoring algorithms from time to time might be indicated. In the negotiation to get credit, as in all negotiations, counterparties should avoid being too predictable.

Another approach, not used here, is to simply refuse to extend credit to applicants caught in a lie. While at first appealing, this is actually problematic, as it requires careful consideration of the difference between a simple error and a lie.

6.2.1 Further research directions

We can extend this research using real data to show that it is profitable to reduce lies in borrower's attributes. In real-world settings, the historical dataset used to build models contains the borrower's characteristics, the repayment performance of the borrower and also a binary variable showing whether the borrower defaulted or repaid their loan. It does not contain any

variable describing whether the borrowers purposely reported falsified attribute values. That is to say, the dataset did not directly tell us whether the borrower lied. Therefore, we have to first consider a method that distinguishes liars and honest borrowers. Furthermore, we understand that using more borrower characteristics may increase the performance of the classification model, if these new characteristics can be measured accurately enough [1]. We can enhance our research by building a multivariate model which takes into account characteristics that predict the repayment behavior of borrowers. To apply discriminant analysis, we must find the joint distribution of all the characteristics used in the model. In addition, our thesis assumed the interest rate charged by the bank to be the same amongst all borrowers. In real-world settings, loan amounts and interest rate will be set according to the credit quality of the borrower. We did model, in a very stylized way, the complicated connection between attributes and loan amounts: but this question deserves a much more methodical study.

As noted in this thesis, we only consider adding lies to the data using a Bernoulli random variable. Furthermore we suppose that all customers who choose to lie will do so by adding the same constant amount to their attributes. However, we can extend our studies by considering different ways to add lies to the data. For example, we can conditionally add lies to the data by looking at the value of the original data, and also set the amount of lies different for each customer by introducing another random variable which determines the strength of lies. In addition, it would be interesting to study how robust these results are in a model in which good borrowers as well as bad ones lie, perhaps for reasons external to the loan-granting process. Another alternative might be to attempt not to correct the attributes stated on forms as modeled here but to identify liars and simply reject their applications.

Bibliography

[1] M. Chong, and M. Davison, Larger Datasets Lead to More Inaccurate Credit Scoring, *Proceedings 59th ISI World Statistics Congress*, Hong Kong, 25-30 August 2013.

[2] M. Chong, C. Bravo and M. Davison, How Much Effort Should Be Used to Detect Fraudulent Applications When Engaged in Classifier-Based Lending?, *Intelligent Data Analysis*, vol. 19, no. s1, pp. S87-S101.

Appendix A

Data Simulation

We used simulated data in this thesis to test the accuracy of our models and present results. Essentially unlimited quantities of synthetic data can be created at very low cost. Many researchers use simulated data to build scientific models, so as to have full detail about the behavior of the data which allows them to identify problems easily. Maldonado and Paredes [2] used synthetic data to predict how likely it would have been for a rejected applicant to have repaid their loan if they had in fact received credit. Banasik *et al*. [1] used simulated data to examine the predictive accuracy of credit scoring models that have been built on a sample of only accepted applicants, and to investigate whether econometric sample selection techniques would improve the accuracy of the model.

The statistical software R [3] has been used to generate our data and this appendix provides the procedure on the way to simulate our dataset. We first simulated a Bernoulli random variable GB to distinguish whether the borrower is a good or bad payer. Good borrowers repay their loans on time, while bad borrowers default. The Bernoulli variable GB takes the value one (indicating a good payer) with probability p, and value zero (indicating a bad payer) with probability $1-p$. If GB is one, we simulate from the corresponding 'good' characteristics X_{1G} and X_{2G}, where X_{1G} and X_{2G} are normally distributed with mean μ_{1G} and μ_{2G} and standard deviation σ_{1G} and σ_{2G} respectively. On the other hand, if GB is zero, we simulate from the corresponding 'bad' characteristics X_{1B} and X_{2B}, where X_{1B} and X_{2B} are normally distributed with mean μ_{1B} and μ_{2B}, and standard deviation σ_{1B} and σ_{2B} respectively. Since we believe each individual borrower's characteristics are related, we set the correlation between X_1 and X_2 to be ρ. The procedure for simulating the entire dataset including X_1 and X_2 is described below:

Step 1. GB, where $GB \sim Bern(p)$

Step 2. $X_1 = \mu_1 + \sigma_1 Z_1$, where $Z_1 \sim N(0,1)$

Step 3. $Z_3 = \rho Z_1 + \sqrt{1-\rho^2} Z_2$, where $Z_2 \sim N(0,1)$ and $Z_1 \perp Z_2$

Step 4. $X_2 = \mu_2 + \sigma_2 Z_3$

We repeat steps 1 to 4 for k times to generate a dataset with N borrowers.

Bibliography

[1] J Banasik, J Crook and L Thomas, Sample selection bias in credit scoring models, *Journal of the Operational Research Society*, **54**(8), (2003), 822-832.

[2] S. Maldonado and G. Paredes, A semi-supervised approach for reject inference in credit scoring using SVMs, *Advances in Data Mining. Applications and Theoretical Aspects*, 10th Industrial Conference, ICDM 2010, Berlin, Germany, July 12-14, 2010, pp. 558-571.

[3] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014, `https://www.R-project.org/`.

Appendix B

Lending Club

Traditional lending business can be described as institution-to-people, where a middlemen such as a bank is usually involved. Banks obtained the supply of money from customers' savings account or other financial instruments such as the selling of bonds and stocks, and on the other hand lend it out to a group of customers at a higher interest rate. Borrowers will then receive their loans out of the money pool that the bank has available. Social lending can be described as people-to-people or peer-to-peer (P2P) lending. The advance development on information systems, e-commerce tools and high speed internet enhance the growth of P2P lending. It is an emerging alternative online lending platform, where individuals lend and borrow money using an online trading system without the help of official financial institutions. It connects people with money to invest to people needing money. On one end, lenders can choose the borrowers to whom she wants to lend money, and on the other end, borrowers can also choose the lenders who provide them with the best interest rate and repayment terms.

Lending Club is a U.S. P2P lending company and was the first P2P lender to register its offerings as securities with the Securities and Exchange Commission. It has the world's largest peer-to-peer lending platform and had been originating over $15.98 BN loans as of December 2015. Lending Club enables borrowers to create unsecured personal loans between $1,000 to $35,000, with a standard repayment period of three years. It provides an online trading platform that connects borrowers to lenders, facilitates a series of transactions from transferring money to borrowers account, collecting payments, imposing penalties, and up to handling default through collection processes. Investors can search through the list of loan request on Lending Clubs website to select loans that they wanted to invest and make money from interest payments. Lending Club on the other hand makes money by charging borrowers an origination fee and investors a service fee.

There are a lot of popular P2P lending platforms in todays world: U.S.- based Prosper [1]

[1] http://www.prosper.com

and Lending Club Corporate, UK-based Zopa Limited [2], and Germany-based Smava GmbH [3]. The convenient use of these P2P lending platforms established an alternative investment instrument, which allows investors to earn higher returns compared to savings and investment products offered by banks. With the continued growth of computer systems and information technology, P2P lending is expected to further develop in the future.

[2] http://www.zopa.com
[3] http://www.smava.de.

Appendix C

Glossary of Terms

Wikipedia [1] and Investopedia [2] define the following common loan and banking terms:

Borrower A person that has applied, met specific requirements, and received a monetary loan from a lender. The individual initiating the request signs a promissory note agreeing to pay the lien holder back during a specified timeframe for the entire loan amount plus any additional fees. The borrower is legally responsible for repayment of the loan and is subject to any penalties for not repaying the loan back based on the lending terms agreed upon.

Creditor A creditor is an entity (person or institution) that extends credit by giving another entity permission to borrow money intended to be repaid in the future. A business who provides supplies or services to a company or an individual and does not demand payment immediately is also considered a creditor, based on the fact that the client owes the business money for services already rendered.

Debtor A debtor is a company or individual who owes money. If the debt is in the form of a loan from a financial institution, the debtor is referred to as a borrower, and if the debt is in the form of securities, such as bonds, the debtor is referred to as an issuer. Legally, someone who files a voluntary petition to declare bankruptcy is also considered a debtor.

Debt Debt is an amount of money borrowed by one party from another. Debt is used by many corporations and individuals as a method of making large purchases that they could not afford under normal circumstances. A debt arrangement gives the borrowing party permission to borrow money under the condition that it is to be paid back at a later date, usually with interest.

Lender A lender is an individual, a public group, a private group or a financial institution that makes funds available to another with the expectation that the funds will be repaid, in addition to any interest and/or fees, either in increments (as in a monthly mortgage payment) or as a lump sum.

Fraud In law, fraud is deliberate deception to secure unfair or unlawful gain, or to deprive

a victim of a legal right. Fraud itself can be a civil wrong (i.e., a fraud victim may sue the fraud perpetrator to avoid the fraud and/or recover monetary compensation), a criminal wrong (i.e., a fraud perpetrator may be prosecuted and imprisoned by governmental authorities) or it may cause no loss of money, property or legal right but still be an element of another civil or criminal wrong. The purpose of fraud may be monetary gain or other benefits, such as obtaining a driver's license or qualifying for a mortgage by way of false statements.

Bibliography

[1] Wikipedia, https://www.wikipedia.org, accessed Dec 3, 2016.

[2] Investopedia, http://www.investopedia.com, accessed Dec 3, 2016.

Curriculum Vitae

Mimi Chong

Education

Doctor of Philosophy, Financial Modeling **2010 - Present**

The University of Western Ontario, London, ON

- Thesis Topic: Game Theory Enhanced Credit Scoring Models
- Secondary Field: Queuing Models in High Frequency Trading
- Received Western Research Scholarship of CAD 22,000 per year

Master of Science, Statistics **2007 - 2008**

University of Toronto, Toronto, ON

- Received Entrance Scholarship of CAD 17,000 for 1 year masters studies

Bachelor of Science [Graduated with Distinction],

Honors Specialization in Mathematical Applications in Economics and Finance, Major in Actuarial Science and Minor in Statistics **2003 - 2007**

University of Toronto, Toronto, ON

- Received Dean's List Scholar of High Academic Achievement

Honors and Awards

Winner, Data Analytics Problem Solving (DAPS) Trophy Competition **Feb. 2013**

The University of Western Ontario

- Participated as part of a seven member interdisciplinary team in an industrial problem solving workshop to innovate new ideas
- Led the statistical analytics part of the project to deal with obstacles that affect real companies

Working Experience

CitiFinancial Canada **Aug 2014- July 2016**

Modeling/Scoring/Analysis Analyst:

- Develop statistical models to forecast gross credit loss towards the effect of macroeconomic factors
- Document the development process and the methodology used in the statistical model
- Back-test and validate scoring models to ensure the accuracy of the model
- Examine the impact of low oil price towards the performance of the business
- Monitor credit bureau reporting process to ensure a low rejection rate on trade-line records

Working Experience (Teaching Assistantships)

The University of Western Ontario **2010 - Present**

Teaching Assistant For: Financial Modeling, Stochastic Processes with Apps in Finance (Graduate Level), Probability and Statistics, Statistical computing and Introduction to Biostatistics

The Hong Kong Polytechnic University **2009 - 2010**

Teaching Assistant For: Engineering Calculus, Linear Algebra, Risk Management and Introductory Statistics [Nominated for the best teaching assistant award]

University of Toronto **2007 - 2008**

Teaching Assistant For: Mathematics of Investment and Credit and Mathematics of Finance

Publications

- Mimi Chong & Matt Davison (2013). "Larger Datasets Lead to More Inaccurate Credit Scoring", *The International Statistical Institute*, (Section CPS009), 3429-3434.
- Mimi Chong, Cristián Bravo & Matt Davison (2015). "How Much Effort Should Be Used to Detect Fraudulent Applications When Engaged in Classifier-Based Lending?", *Intelligent Data Analysis*, vol. 19, no. s1, pp. S87-S101.

Conference Presentations

- **Business Analytics in Finance and Industry, Santiago, Chile** **2014**
  Paper Presented: "How much Effort Should Be Spent to Detect Fraudulent Applications When Engaged in Classifier-Based Lending?" (co-authored with Matt Davison and Cristián Bravo).
- **Big Data Synergy @ Western, London ON., Canada** **2014**
  Poster Presented: "Classifiers On Big Data to Detect Credit Fraud" (co-authored with Matt Davison).
- **59th ISI World Statistics Congress** **2013**
  Paper Presented: "Larger Datasets Lead to More Inaccurate Credit Scoring" (co-authored with Matt Davison).