


Article

A Method for Visualizing Posterior Probit Model Uncertainty in the Early Prediction of Fraud for Sustainability Development

Shih-Hsien Tseng ^{1,*} and Tien Son Nguyen ² 

¹ Department of Industrial Management, National Taiwan University of Science and Technology, 43 Sec. 4 Keelung Road, Daan District, Taipei 106335, Taiwan

² Institute of Industrial Management, National Central University, 300 Zhongda Road, Zhongli District, Taoyuan City 32001, Taiwan; tonymfu16@gmail.com

* Correspondence: shtseng@mail.ntust.edu.tw

Abstract: Corporate fraud is not only curtailed investors' rights and privileges but also disrupts the overall market economy. For this reason, the formulation of a model that could help detect any unusual market fluctuations would be essential for investors. Thus, we propose an early warning system for predicting fraud associated with financial statements based on the Bayesian probit model while examining historical data from 1999 to 2017 with 327 businesses in Taiwan to create a visual method to aid in decision making. In this study, we utilize a parametric estimation via the Markov Chain Monte Carlo (MCMC). The result show that it can reduce over or under-confidence within the decision-making process when standard logistic regression is utilized. In addition, the Bayesian probit model in this study is found to offer more accurate calculations and not only represent the prediction value of the responses but also possible ranges of these responses via a simple plot.

Keywords: financial statement fraud; bayesian probit model; standard logistic regression; Markov Chain Monte Carlo

MSC: 62F15; 65C40



Citation: Tseng, S.-H.; Nguyen, T.S. A Method for Visualizing Posterior Probit Model Uncertainty in the Early Prediction of Fraud for Sustainability Development. *Axioms* **2021**, *10*, 178. <https://doi.org/10.3390/axioms10030178>

Academic Editor: Hari Mohan Srivastava

Received: 20 June 2021
Accepted: 1 August 2021
Published: 4 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the last few decades, many senior managers have been caught using phony financial statements to cheat stakeholders or manipulate stock prices in an attempt to funnel profits. As such, corporate fraud has long been a serious problem, particularly when it involves financial statements. Ironically, the information contained in these documents has remained as one of the key indicators that fraud has taken place [1,2]. Fraudulent activities have not only directly resulted in significant losses for stakeholders and severe punishments for the accounting institutions involved, but they have also significantly altered trading practices in the financial market. According to the Association of Certified Fraud Examiners (ACFE) in 2020, a total of 2504 cases from all over the world with an average loss of 5% revenue was due to corporate fraud which is equivalent to the loss of Gross World Product (GWP) about USD 3.6 trillion [3]. Although it is possible to detect corporate fraud, the ACEF still holds that it is indeed ubiquitous, and that no organization can be completely immune from this threat. The complex causes of fraud are explained in the agency problem theory, earning management, fraud triangle, and the GONE theory [4]. According to the fiduciary norm, managers must act solely in the interests of the principal, neglecting all others [5]. If the principal and agent are at odds, the latter will tend to focus on his or her interests which has attracted much attention over the years. Moreover, Song, et al. [6] have voiced concern over the privatization of many state-owned businesses in China, which may be problematic because, previously, the interests of state-owned companies have aligned with those of the nation. However, with privatization comes market-oriented goals, meaning effective performance and profit become the primary objectives for the corporation.

Although earnings management in and of itself is a legitimate practice, corporate managers often manipulate it for their benefit. For example, they interfere with financial reporting when preparing statements to purposely mislead readers [7] and the public about the performance of specific enterprises [8], which may cause investors to make misinformed decisions. In this way, unethical earnings management may lead to fraudulent financial reporting. The fraud triangle theory was firstly proposed by Cressey [9], which effectively explains various aspects of this crime and has become the foundation of SAS No. 99. Many scholars have applied this theory in a variety of ways [10]. For instance, Brennan and McGrath [11] argue that the most common form of this crime is the creation of fraudulent transaction records to meet expected profit levels. According to Skousen et al. [12], the rapid growth of company assets, the need for cash, and an increase of external capital are frequent indicators that fraud is taking place. In a study of 64 British firms accused of fraud, Hollow [13] discovered that financial pressure plays a critical role in this crime. Bologna, Lindquist, and Wells [4] proposed the GONE theory, which posits that greed, opportunity, need, and exposure is closely associated with fraud. From this perspective, it becomes obvious that greed and need are personal factors while opportunity and exposure are environmental or systemic. Although all four features must be present for fraud to occur, they do not need to exist simultaneously.

Altman [14] proposed the following five financial ratios to effectively determine such difficulties: (1) the ratios of current assets/total assets, (2) retained earnings/total assets, (3) earnings before interest and taxes (EBIT)/total assets, (4) equity value/total liabilities, and (5) total sale value/total assets. This model has also been used to predict the bankruptcy risk of various companies. The lower the company scores, the higher the possibility of bankruptcy. Likewise, if a company is in grave danger, the manager will be under extreme financial pressure. According to Cressy's fraud triangle theory, the stress of this kind tends to promote corporate fraud. Also, Persons [1] found that financial leverage, asset turnover rate, asset portfolio, and the size of the company are closely associated with fraudulent financial reporting. Thus, managers of smaller companies with high levels of financial leverage and low asset turnover rates will be most likely to commit financial reporting crimes. The findings from many empirical studies indicate that the type of corporate governance greatly impacts the behavior of managers and the company's overall performance [15–17]. Xie, et al. [18] hold that large boards will be more likely to be comprised of experts with a variety of backgrounds and areas of specialization, who will be able to contribute to the effective supervision of managers and, thus, mitigate the agency problem. According to Beasley [15], if the board has a high percentage of external or independent directors who have extended terms of office, and the company has a significant number of external shareholders, the possibility of fraud will be greatly reduced.

The logistic regression model has long been studied for academic studies of fraud, and it remains the prevailing technique for studying this devastating crime. However, many scholars have attempted to include various perspectives to improve its flexibility. Specifically, they included related variables such as conventional financial indicators [1], audit quality [19,20], corporate governance [15–17], and the principle of stability [21,22] in the fitted models. Lin [23] integrated the principles of conventional financial indicators, corporate governance, and stability into the fitted model and showed better performance than the considered conventional financial indicators and the corporate governance factor within the model. Ensemble modeling techniques have become increasingly popular to enhance classification accuracy [24–26]. Recently, Tseng et al. [27] employed these methods to investigate the impact of bias, multicollinearity, and erroneous input patterns on model analysis. In other words, if parameter uncertainty in the fitted model is not considered, this oversight might easily lead to erroneous inferences and flawed estimates of quality. The purpose of this study is to take this critical element into account for generating multiple models from the posterior distribution through Bayesian probit modeling via Markov Chain Monte Carlo (MCMC). We implemented the MCMC method for developing relatively realistic predictive models largely from the posterior distribution even in the absence of

closed-form parameters. Also, we show the visual distribution of the prediction values for better understanding the results under comparison of two models. In addition, we construct 13 indicators of corporate governance and provide a financial overview to improve fraud detection. This study is motivated by the following questions: (1) How to visualize the effects of uncertainty bias for better decisions? How to handle the overestimation or underestimation of statistical models? (3) Which method can enhance the predictive power and ameliorate the effects of model uncertainty? To answer these questions, we aim to reduce the bias of parametric estimation based on the Bayesian probit model and compare it with the standard logistic model through visualization. This study is organized as follows: the review of related studies, analysis of the causes of fraud, and a description of the predictive model are shown in Section 2. In Section 3, we present the structure of the model and define the variables. Section 4 provides an in-depth discussion of the data gathering process, parametric estimation, and analysis of the results while Section 5 includes the conclusion and recommendations for future research.

2. Literature Review

2.1. Related Studies of Fraud Detection

Fraud detection has been studied for a long time, with many techniques and models such as logistic models, decision trees, artificial neural networks [28–30], support vector machines [31], and random forests [32] or data engineering methods [33] which have proven to be quite precise. The most famous model is the Z-score, which is commonly used even nowadays for predicting financial distress and fraud [34]. Summers and Sweeney [35] used the logit model to study 51 companies that were under investigation by The Wall Street Journal for financial statement fraud from 1980 to 1987. The researchers matched the samples from the same number of no-fault companies following the standard industry classification code (SIC code). They found that company insiders who commit fraud tend to sell significant numbers of shares in order to reduce the quantity available for others to buy, which obviously also reduces the percentage of shares held by the company. Imhoff [36] suggests that substantive change is necessary to improve corporate governance. Problems with accounting or auditing procedures will not be solved until boards are given sufficient information to operate independently and are allowed to act on behalf of the shareholders.

In practice, fraudulent financial reporting is associated with managers who can easily override or change the internal control procedures while appearing to be loyal to the company [15]. Under these circumstances, managers can easily manipulate earnings and present falsified financial reports. Desai [37] suggests that many corporate scandals have been caused by the exaggeration of profits. Managers tend to report gross profits in the capital market and taxable profits to governmental agencies to avoid paying taxes, which leads to the creation of fraudulent financial statements. Davidson, et al. [38] studied the effect of corporate governance on earnings management by analyzing 434 companies listed in the exchange. They discovered that most non-executive directors on the board and audit committees would be less likely to manipulate earnings if the board is independent. Perols and Lougee [39] argue that managers engaged in acts of fraud begin to manipulate earnings a few years before the crime is detected. The level of adjustment may even exceed that of predicted growth, or they may exaggerate their revenues to commit financial statement fraud.

Many researchers also suggest that the quality of audits can be guaranteed [19], and fraud will much less likely [40] if financial statements are audited by large accounting firms. Although this theory is not directly observable, Hribar, Kravet, and Wilson [20] who used accounting fees as a surrogate variable, found that the size of the fees may reflect the level of reliability of the statements. Kamarudin, Ismail, and Mustapha [22] have a different perspective on this controversial issue. After analyzing data from 184 companies from 2003 to 2010, they found that most that were guilty of fraud tended to practice “aggressive accounting” including claiming revenue prematurely or over-optimism and the timely identification of loss. Although these practices are not against the law, they are considered

negligent because their presence indicates that financial statements must be compiled a second time which calls into question the reliability and quality of financial reporting. In recent years, data mining and machine learning techniques have shown many advantages to traditional statistical tools in fraud detection, but we are still trying to explain this “black box” to reduce the bias of models [26]. Perols [28] found that logistic regression outperforms neural networks and decision trees. Furthermore, the Bayesian Belief Network model outperforms decision trees and neural network models for identifying fraudulent financial statements, and it also can utilize ten-fold stratified cross-validation [30]. Also, many scholars and practitioners prefer the Bayesian methods rather than machine learning or deep learning models due to the limit of data and lack of interpretability [41]. After analyzing the development of fraud detection models, it becomes clear that the accuracy of models depends heavily on gauging financial indicators.

2.2. Comparing the Bayesian Probit Model to the Standard Logistic Model

Over the years, many scholars have performed logistic analysis using the Bayesian model [42–45] to correct parametric estimation errors and establish a more realistic model. This method has been extensively applied to various domains of research. Gerlach, et al. [46] applied the Bayesian probit model to 63 items within financial statements and used step-wise regression to select appropriate variables to create a logistic model specifically for forecasting changes in corporate earnings. Lately, the Bayesian probit model, which is widely used in the domain of statistics, has attracted much attention in the field of social science [47]. In a similar vein, Rossi, et al. [48] adopted this model to analyze many marketing problems and help managers make more informed decisions. The difference between the Bayesian probit model and the standard logistic model is that the estimation of parameters under the latter is based on the Maximum Likelihood Estimation (MLE). This iterative method of calculation is necessary for determining non-linear solutions, which causes the expression of parameters to be in closed-form. After calculating the coefficient, the chi-square can be used to test its significance. Another common method is the Wald test, which conforms to the standard normal distribution with a null hypothesis [49].

Although some researchers argue that the most effective sample size for the standard logistic model is only ten or more [50], the process of mathematical inference requires a larger sample size that is substantial enough to effectively approximate the chi-square or normal distribution. However, the prior assertion cannot be ignored. Researchers always use a sample size of less than 100 for corporate fraud studies due to the prolonged time it takes to reach verdicts in such cases. For this reason, there are not enough types of samples to conduct a valid study. These limited sample sizes remain one of the inherent shortcomings of the standard logistic model. In addition, that model operates through the pairing of samples. The common ratio of pairing companies that have been accused of fraudulent activities with no-fault companies is 1:1 or 1:2. In reality, it would be difficult to find two companies of similar size in the same industry. For example, in an oligarchic market, the size of companies varies significantly. At this point, because it is so difficult to find companies in good standing to use for analysis, the results of this study would be somewhat biased. This is yet another shortcoming of the logistic model. Although it is unnecessary to assume that the independent variables are from the normal distribution, after model fitting and computing the confidence interval between the independent and the dependent variable, the standard normal distribution method of the Wald test is required. Therefore, this model may not be stable enough to detect fraud, which is a third shortcoming of the standard logistic model. Whether or not the results from this model can effectively map the relationship between the variables, is another issue to be explored in the future.

Due to these shortcomings, we adopted the Bayesian probit model in conjunction with the MCMC for this study to overcome the aforementioned constraints [51]. After utilizing simulation to redistribute the parameters, we compared the posterior probability to the prior probability via the Bayesian probit model to create a realistic scenario. This model is

also more effective and stable than others for determining early signs of fraud. In summary, this model can help to effectively eliminate the bias of parametric estimation. However, it has not been popularly applied by researchers in particular of financial statement fraud. Thus, the objective of this study is an attempt to use the Bayesian probit model to more effectively analyze financial statement fraud and compare it to the results of the standard logistic model to provide a more accurate reference and decision-making guide.

3. Methods

3.1. Notations

In the Bayesian probit model, we noted y that represents corporate fraud as 1, while all others were noted as 0. Therefore, the equation for determining the probability of fraud is $F(x_i; \beta) = P(y_i = 1|x_i; \beta)$, and non-fraud is $P(y_i = 0|x_i; \beta) = 1 - F(x_i; \beta)$. As such, the logistic function $g(x_i)$ is also referred to as an odds ratio, as expressed in the equation below:

$$g(x_i) = \ln \frac{F(x_i|\beta)}{1 - F(x_i|\beta)} = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i \tag{1}$$

where $i = 1, 2, \dots, n$. that represents the sample size of the model; $j = 1, 2, \dots, p$. symbolizes the individual variables; $F(x_i; \beta)$ is the probability of fraud while ε_i represents the residual effects.

For the logistic function, parameter β was calculated via MLE, and the i^{th} term likelihood function was determined as $l_i(\beta) = F(x_i|\beta)^{y_i} [1 - F(x_i|\beta)]^{1-y_i}$, which could be expanded into Equation (2). I assumed that each variable was independent, and that the likelihood function of the model would be the product of all items, as shown in Equation (3). According to the Bayesian inference, the posterior probability would be directly proportional to the product of the likelihood function and prior probability, which is shown in Equation (4).

$$l_i(\beta) = \left(\frac{e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}}} \right)^{y_i} \left(1 - \frac{e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}}} \right)^{(1-y_i)} \tag{2}$$

$$l(\beta) = \prod_{i=1}^n \left[\left(\frac{e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}}} \right)^{y_i} \left(1 - \frac{e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}}} \right)^{(1-y_i)} \right] \tag{3}$$

$$P(\beta|Y, X) = \frac{P(Y, X|\beta)P(\beta)}{P(Y, X)} \propto \text{Likelihood} \times \text{prior} \tag{4}$$

Furthermore, we summarize the sequence of the proposed method as a flowchart in Figure 1.

3.2. MCMC Parameter Estimation

There has recently been a resurgence in the use of Bayesian regression methods, in part due to the popularity of the MCMC approach [48]. In this study, our model was derived from a combination of the Markov Chain and the Monte Carlo methodologies. Based on random sampling from the Markov Chain, the Monte Carlo method is used to estimate the integration of problems that have no analytical solutions or to analyze difficult and complicated probability distributions.

When employing the Markov Chain, we assumed that if $\beta^0, \beta^1, \beta^2, \dots$ are a series of random variables, then β^{t+1} would be generated from the conditional probability of $P(\beta^{t+1}|\beta^t)$, and its value would only depend on β^t and would not be related to $\{\beta^0, \beta^1, \beta^2, \dots, \beta^t, \beta^{t-1}\}$. When time t increases, the distribution would become stationary and independent from t and β^0 . However, if the probability could not fit into a standard distribution, we would need to apply the Monte Carlo method to obtain an accurate estimation. For instance, if β is the random variable of the model parameter, and we assume that it conforms to the posterior probability distribution $\pi(\beta)$, then $f(\beta)$ would be the

expected value of the probability distribution, as shown in Equation (5). Sometimes, if it is too difficult or even impossible to calculate the integration using Equation (5), we employ the Monte Carlo integration, which is based on random sampling from $\pi(\beta)$ for selecting $\{\beta^1, \beta^2, \dots, \beta^m\}$ and can be used to accurately estimate the mean value of the samples to approximate the expected value of the probability distribution $f(\beta)$. The process is shown in Equations (5) and (6):

$$E[f(\beta)] = \int f(\beta)\pi(\beta)d\beta \tag{5}$$

$$E[f(\beta)] \approx \frac{1}{m} \sum_{t=1}^m f(\beta^t) \tag{6}$$

where β^t represents the t^{th} sampling result when $t \geq 0$.

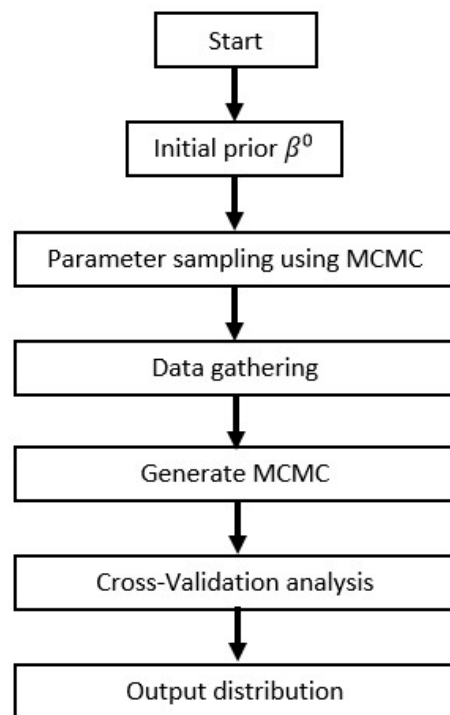


Figure 1. The sequence of the proposed method.

It becomes clear that if the initial value were different, the average estimation result would also change. Thus, if we could establish that $\phi() = \pi()$, we could ignore the burn-in sample of the previous r^{th} test, utilize the sampling result with interval k , and solve the above problem via Equation (7).

$$E[f(\beta)] = \lim_{m \rightarrow \infty} \frac{1}{m - r} \sum_{t=r+1}^m f(\beta^t) \tag{7}$$

In this study, we applied the Gibbs sampling method (an MCMC algorithm), a special type of the Metropolis-Hastings algorithm proposed by [52] to obtain the following observations. According to this method, we determined the result of the i^{th} sampling of $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ from the m th sampling as $\beta^i = (\beta_0^i, \beta_1^i, \dots, \beta_p^i)$ by following the three steps shown below.

Step 1: We found the initial value of $\beta^0 = (\beta_0^0, \beta_1^0, \dots, \beta_p^0)$ of a given parameter and set the sampling frequency to m .

Step 2: We conducted an $i^{th} + 1$ sampling to determine the value of $\beta^{i+1} = (\beta_0^{i+1}, \beta_1^{i+1}, \dots, \beta_p^{i+1})$ and updated the value for each instance, as shown in Equation (8).

$$\begin{aligned} \beta_0^{i+1} &\sim \phi_0(\beta_0 | \beta_1^i, \beta_2^i, \dots, \beta_p^i, Y, X) \\ \beta_1^{i+1} &\sim \phi_1(\beta_1 | \beta_0^{i+1}, \beta_2^i, \dots, \beta_p^i, Y, X) \\ &\vdots \\ \beta_{p-1}^{i+1} &\sim \phi_{p-1}(\beta_{p-1} | \beta_0^{i+1}, \beta_1^{i+1}, \dots, \beta_p^i, Y, X) \\ \beta_p^{i+1} &\sim \phi_p(\beta_p | \beta_0^{i+1}, \beta_1^{i+1}, \dots, \beta_{p-1}^{i+1}, Y, X) \end{aligned} \tag{8}$$

Step 3: We used the parametric values from the sampling to repeat step 2 until we reached the end of the m^{th} sample.

After estimating via the Gibbs sampling, in order to verify that the Markov Chain reached stationarity, we used the Autocorrelation Function (ACF) to monitor the convergence of the chain [48,53]. Then, we selected the number series $\{\beta^m : m = 0, 1, 2, \dots\}$ from the m value of the Markov Chain. When m approximated infinity, β^m changed to β . At this point, β was the random variable from the joint probability distribution, $f(\beta)$ and we accomplished our estimation goal.

3.3. Creating the Fraud Detection Model

During the data-gathering phase, n represents the total number of companies and X_i signifies all the predictive variables of the i th company. These could include continuous or dispersed variables, such as financial indicators, corporate governance variables, principles of stability, and the size of the company, which will be explained in detail in Section 3.4. In the model, if $y_i = 1$, this would indicate that an act of fraud had taken place at i^{th} company. If $y_i = 0$, this would suggest that employees at i^{th} company were innocent of this crime. In this study, my analysis was based on the binary probit model language of the R statistical software for sampling and estimation, as shown in Equation (9).

$$\begin{cases} y_i = 1 \text{ if } z_t \geq 0 \\ y_i = 0 \text{ if } z_t < 0 \end{cases} \quad \forall i = 1, 2, \dots, n, \forall t = 0, 1, 2, \dots, p \tag{9}$$

$$z_i = X_i\beta_t + \varepsilon_i, \varepsilon_i \sim N(0, 1)$$

where $Y_i = (y_1, y_2, \dots, y_n)$ is a vector of $n \times 1$ which is used to determine if employees at the i^{th} company which is engaged in fraud. $Z_i = (z_1, z_2, \dots, z_n)$ is also a vector of $n \times 1$ and the aggregate of the continuous potential variables that correspond to Y_i . As such, the model structure that corresponds to the i^{th} company is shown in Equation (10).

$$X_{i,t} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}_{n \times (p+1)}, \beta_t = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1) \times 1}, \varepsilon_i = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n \times 1} \tag{10}$$

In this model, the cutoff point of value in the judgment of $\{Y_i\}$ differs from that in the logistic model. Thus, before we could begin any analysis, we converted the scope covered by $\{X_i\}$ to a range within the closed-form of $[-1, 1]$ [48], as shown in Equation (11).

$$\{X_i\} = \frac{\text{original}\{X_i\} - \frac{\text{Max}(\text{original}\{X_i\}) + \text{Min}(\text{original}\{X_i\})}{2}}{\frac{\text{Max}(\text{original}\{X_i\}) - \text{Min}(\text{original}\{X_i\})}{2}} \tag{11}$$

In the fraudulent financial statement prediction model proposed in this study, the only observed values were $\{X_i\}$ and $\{Y_i\}$. The estimation parameters were the aggregate of β in the multiple of $p + 1$ denoted as $\{\beta_t\} = (\beta_0, \beta_1, \dots, \beta_p)$ while the posterior

probability $\{\beta_t\}$ featured the closed-form parameters. As such, I used the Gibbs sampling of the posterior probability distribution to estimate the joint probability distribution of $f(\beta)$ of $\{\beta_t\}$.

3.4. Description of Variables

In this study, the fitted model has constructed 14 variables that are similar to Lin [23]. The operational definitions are discussed below:

- Dependent Variables:

We used binary classification to categorize the variables in this equation. The fraudulent company was noted as 1 and the no-fault company was 0.

- Independent Variables:

In this study, there were 13 independent variables from the following categories: the “five financial ratios,” proposed by Bernstein [54], included profitability, liquidity, growth, utility, and financial structure (Table 1), corporate governance variables (Table 2), and conservative accounting variables.

Table 1. Independent Variables from Bernstein and Wild Bernstein [54].

| 5 Financial Ratios | Equation | Index | Index Equation |
|-----------------------------|--|---|--|
| Profitability (β_1) | Revenue growth ratio | Revenue growth ratio | (Net income of T period–Net income of T-1 period)/(Net income of T-1 period) |
| Liquidity (β_2) | (Current ratio + Working capital ratio)/2 | Current ratio | Current assets/Current liabilities |
| | | Working capital ratio | (current assets—current liabilities)/Total assets |
| Growth (β_3) | (Ratio of return on assets + Net profit rate + Net operating profit ratio)/3 | Return on assets ratio | Income after taxes/Total assets |
| | | Net profit ratio | Income after taxes/Sales revenue |
| | | Net operating profit ratio | Net operating income/Sales revenue |
| Utility (β_4) | (Accounts receivable to total assets ratio + Sales to total assets ratio)/2 | Accounts receivable to total assets ratio | Accounts receivable/Total assets |
| | | Sales to total assets ratio | Sales revenue/Total assets |
| Structure (β_5) | Debt ratio + Net liabilities ratio)/2 | Debt ratio | Total liabilities/Total assets |
| | | Equity Ratio | Total liabilities/Shareholders’ equity |

Table 2. Independent Variables from Corporate Governance.

| Corporate Governance Variable | Equation/Explanation |
|--|--|
| Number of board members (β_6) | Number of directors |
| Ratio of external directors (β_7) | The ratio of the number of external directors to total director’s seats |
| The chairman also holds the position of general manager (β_8) | Dummy variable, chairman who also holds the position of general manager is represented by 1. If not, it is represented by 0. |
| Percentage of shareholding by directors (β_9) | The quantity of shares held by the directors/Total outstanding shares at the end of the period. |
| Percentage of shareholding by institutional investors (β_{10}) | The ratio of institutional investors in the company. |
| Deviation between one’s voting rights and earnings (β_{11}) | Voting rights minus earnings distribution rights |

We adopted Givoly and Hayn [21] hypothesis of stable variables, which states that the greater the Conservative Accounting (CONACC) value, the more conservative the accounting policy of the company.

$$CONACC(\beta_{12}) = -\frac{1}{3} \sum_{t=-2}^0 \frac{(\text{Earnings before extraordinary items} + \text{depreciation} - \text{cash flow from the operation})}{\text{Total assets at the beginning of study timeframe}}$$

- Control Variables

Size of the company $\beta_{13} = \ln(\text{Asset Size})$.

4. Bayesian Modeling

4.1. Sample Data

In this section, we applied the data organization as Lin [23] for adapting the framework and utilizing the MCMC method to thoroughly analyze. The income is chosen before extraordinary gain (loss). However, since enterprises in Taiwan have already adopted the IFRS accounting standards, income (loss) for continuing is more appropriate than before.

$TA(\beta_{12}) = [\text{income (loss) for continuing} + \text{depreciation} - \text{cash flow from operations}] / \text{average total assets}$:

$$CONACC = -\frac{1}{3} \sum_{t=-2}^0 TA \tag{12}$$

We analyzed companies that had been convicted of fraud in a court of law for crimes such as insider trading, stock price manipulation, and fraudulent financial statements between 1999 and 2017. The reason we used the dataset until 2017 was because most of the recent investigations could not be completed yet. Of the 327 companies investigated, 109 were found guilty. The 1:2 ratio method was used to match them with 218 companies that had not engaged in fraud (see Table 3).

Table 3. Description of fraud samples.

| | | | |
|--------------------------------|--|---------------------------------|----|
| Definition of Fraud | According to Statements on Auditing Standards (SAS) No. 43: One or more managers, those in governance, or employee level personnel have deliberately used deception to obtain improper or illegal gains. | | |
| Fraud Sample Screening Methods | Announcements by the Securities and Futures Investors Protection Center Court Judgments | | |
| Fraud Sample Years | 1999–2017 | | |
| Fraud Sample Types | Type 1 | Stock Price Manipulation | 42 |
| | Type 2 | Falsifying Financial Statements | 32 |
| | Type 3 | Insider Trading | 35 |
| | Total 109 | | |

Moreover, 109 companies that had engaged in fraud spanned a total of 35 different industries. Although the crimes covered a wide range of industry categories, they did not all include special financial statement layout items such as the financial industry, securities, or insurance industries and were very similar in this way. The selection criteria used for pairing companies were based on the industry to which the fraudulent company belonged, and the fact that the asset gaps did not exceed 40% during the same year. The goal is to match two innocent companies with one guilty company of fraud. Corporate information data published by the Taiwan Economic Journal (TEJ) was used in the study. We collected all the data from the year the fraudulent activities took place (T), 1 year prior to the fraudulent activities (T-1), 2 years prior (T-2), and 3 years prior (T-3). Data from 327 enterprises and a total of 981 data items were used to establish the analysis model. The

fraud distribution by industry is shown in Table 4. According to Table 4, a large portion of the fraud detection is from the semiconductor industry with 10.1%, while motherboards stay behind with 7.3%, compared to 35 different industries. In addition, most of the frauds were detected from the 2005–2009 period compared to other periods. Besides, around 30% of industries were detected as fraud with only one company from 1999 to 2017 such as glass ceramics, communication equipment or foods, and animal feed.

Table 4. Distribution of companies engaged in fraud by industry and year.

| | 1999 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | Total | Percentage |
|----------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|-------|------------|
| Hardware and Furniture | | | | | | | 1 | 1 | | | | | | | | | | | 2 | 1.8% |
| Motherboards | | | | | | 1 | 2 | 2 | | 1 | 2 | | | | | | | | 8 | 7.3% |
| Semiconductors | | | | | | 3 | 2 | | 1 | 1 | | 1 | | 1 | | 2 | | | 11 | 10.1% |
| Petrochemicals | | | 1 | | | | 1 | | 1 | | | | | | | | | | 3 | 2.8% |
| Optoelectronics | | | | | | 1 | 1 | 1 | | | | | | 1 | 1 | | 1 | 1 | 7 | 6.4% |
| Garments | | | | | | | 1 | | 1 | | | | | | | | | | 2 | 1.8% |
| Bicycles | | | 1 | | | | | | 1 | | | | | | | | | | 2 | 1.8% |
| Automotive Components | | | | | | 1 | | | | | | | 1 | | | | | 1 | 3 | 2.8% |
| Textiles | | | | | | | | 1 | | | | | 1 | 1 | 1 | | | | 4 | 3.7% |
| Basic Metals | | | | | | 1 | | 1 | 1 | | | | | | | | 1 | | 4 | 3.7% |
| Metal Products | | 1 | | | | | | | | | | | | 1 | | | | | 2 | 1.8% |
| Construction | 1 | 1 | | | | | | 1 | 1 | | | | | | | | 1 | 1 | 6 | 5.5% |
| Glass Ceramics | | | | | | | | | | | 1 | | | | | | | | 1 | 0.9% |
| Ocean Freight | | | | | | | | 1 | | | | | | | | | | | 1 | 0.9% |
| Freight Warehousing | | 1 | | | | | | | | | | | | | | | | | 1 | 0.9% |
| Software Services | | | | | 1 | | | | | 2 | | | | | | 1 | | | 4 | 3.7% |
| Communication equipment | | | | | | | 1 | | | | | | | | | | | | 1 | 0.9% |
| Weaving | | | | | | | | | 1 | | | | | | | | | | 1 | 0.9% |
| Dairy | | 1 | | | | | | | | | | | | | 1 | | | | 2 | 1.8% |
| Information Channels | | | | | 1 | | | 1 | | | | | 1 | | 2 | | | | 5 | 4.6% |
| Electronics Equipment | | | | | | 1 | | | | 1 | | | | 1 | | | | 1 | 4 | 3.7% |
| Electronic Components | | | | | | | 3 | 1 | 1 | 2 | | 1 | | | 2 | 1 | 1 | | 12 | 11% |
| Electrical Wires | | | | 1 | | | | | | | | | | | | | | | 1 | 0.9% |
| Electrical Products | | | | | | | | 1 | | | | | | 1 | | | | | 2 | 1.8% |
| Network Equipment | | | | 1 | | | | | | 1 | | | | | | 1 | | | 3 | 2.8% |
| Shoes and Suitcases | | | 1 | | | | | | | | | | | | | | | | 1 | 0.9% |
| Resin | | | | | | | | | | | 1 | 1 | | | | | | | 2 | 1.8% |
| Machinery Industry | | | | | | | 1 | 1 | 1 | 1 | | | | | | | | | 4 | 3.7% |
| Medical Supplies | | | | | | | | 1 | 1 | | | | | | | | | | 2 | 1.8% |
| Medical Pharmaceuticals | | | | | | 1 | | | | | | | 1 | | | | | | 2 | 1.8% |
| Chemical Material Products | | | | | | | | | | | | | | | 2 | | | | 2 | 1.8% |
| Other Electronics | | | | | | | | | | | | 1 | | | | | | | 1 | 0.9% |
| Tourism and Dining | | | | | | | | | | | | | | | | | 1 | | 1 | 0.9% |
| Foods and Animal Feed | | | | | | | | | | | | | | 1 | | | | | 1 | 0.9% |
| Cement Products | | | | | | | | | | | | 1 | | | | | | | 1 | 0.9% |
| Total | 1 | 4 | 3 | 2 | 2 | 9 | 13 | 13 | 10 | 9 | 4 | 5 | 4 | 7 | 9 | 5 | 5 | 4 | 109 | 100% |

4.2. Prior Distributions

The corresponding probability distributions prior to estimation were assigned to all unknown parameters in the model, including the 14 constant terms. The β prior probability $\bar{\beta}$ in this study was set as the average and the A^{-1} normal distribution of the variances, which were calculated using $\bar{\beta}$ Equation (13) by A^{-1} with $v_0 = 3$ [48].

$$\bar{\beta} = \begin{bmatrix} 0 \\ 0 \\ M \\ 0 \end{bmatrix}_{14 \times 1}, \quad A = v_0 S_X = v_0 \begin{bmatrix} s_1^2 & 0 & \Lambda & 0 \\ 0 & s_2^2 & \Lambda & 0 \\ M & M & O & M \\ 0 & 0 & \Lambda & s_{14}^2 \end{bmatrix}_{14 \times 14} \quad (13)$$

where $S_X = \text{diag}(s_1^2, s_2^2, K, s_{14}^2)$ and $s_j^2 = \frac{\sum_i (x_{ij} - \bar{x}_j)^2}{n-1}$.

4.3. Sampling and Modeling

The parameter of the Bayesian probit model used in the study was estimated according to the MCMC procedure described in the previous chapter. The number of Gibbs samplings was set to 1 million ($R = 1$ million), the sampling interval was 10 (keep = 10), and a total of 100 thousand iterations were obtained. Next, the first 20 thousand sampling results were discarded (burn-in = 20 thousand) and the remaining 80 thousand were determined as the joint probability distribution of the parameters, which were used to calculate the detection capacity and range of the fraud warning model.

K-fold cross-validation was used in this study to establish and analyze the model. The 327 companies were divided into 10 groups according to the three different years using a ratio of 1:2 between fraudulent and non-fraudulent companies. The first nine groups were made up of 33 companies, and the last group contained only 10. I used one as a test group, and the remaining nine were used as training sets. The testing was carried out 10 times, and a different group was chosen to be the test set each time to most efficiently calculate the predictive ability of the model. Besides the first-order term, an interaction term (full second-order) is also added that could represent the analysis results by a particular degree according to Allen and Tseng [55].

4.4. Prediction Results from the Standard Logistic and Bayesian Probit Models

The results of the first-order model are shown in Figures 2–4. Each graph on the box-and-whisker plot was drawn according to the prediction results and was estimated from 80,000 iterations using MCMC. The red dot represents the prediction result of the general logistic model. According to the Cross-Validation result in Figure 2, only Set 4 and Set 8 are stable by using the general logistic model, while others are uncertain in the T-1 period. In the T-2 period, most of the logistic model predictions are stable more than in the T-1 period but the uncertainty seems to increase during the T-3 period. Overall, the figures show that the single result of the logistic model fell within the 80,000 iterations that were estimated using MCMC, which indicates that the logistic model results were quite unstable. However, the MCMC was able to estimate the overall distribution and provided more abundant information.

4.5. Comparison of the One-Time Model and the Interaction Term Model

Moreover, Figures 5–7 illustrate the results of the first order and the interaction term models. Each graph on the box-and-whisker plot was also drawn based on the prediction results from 80,000 iterations that were estimated using the MCMC. The red dot represents the prediction result of the general logistic model. The figures also indicate that this model's results were quite unstable and often produced over- or under-estimations. Furthermore, the predictive accuracy of the interaction term model was generally higher than that of the one-time model.

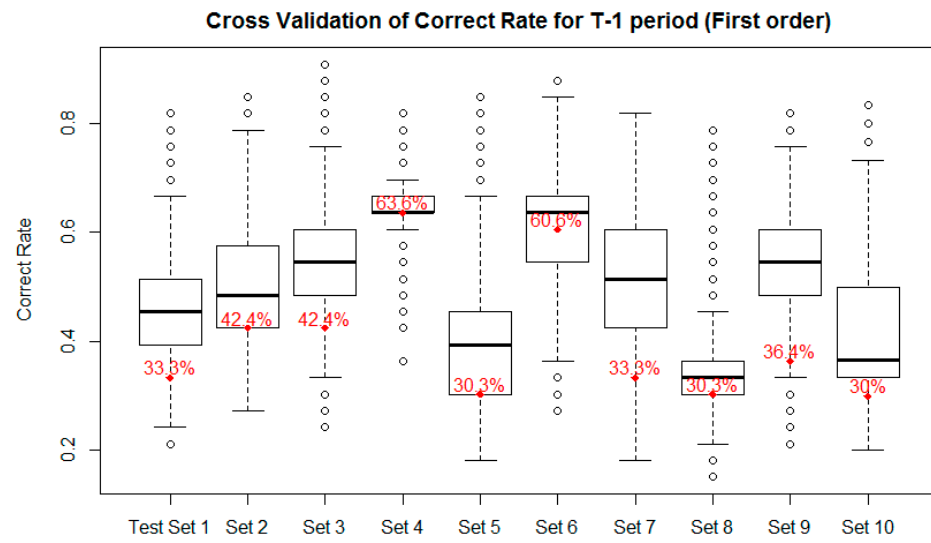


Figure 2. Cross-Validation of Correct Rate for T-1 period (First Order).

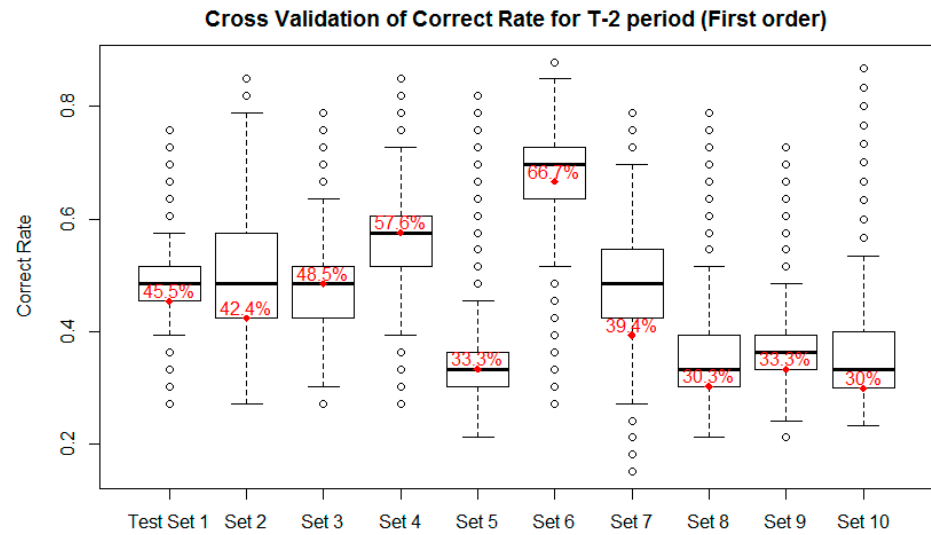


Figure 3. Cross-Validation of Correct Rate for T-2 period (First Order).

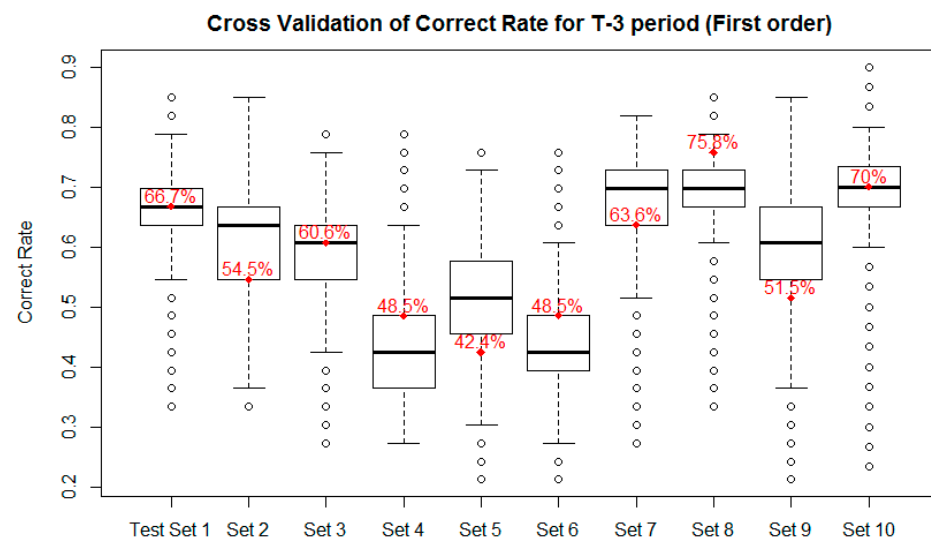


Figure 4. Cross-Validation of Correct Rate for T-3 period (First Order).

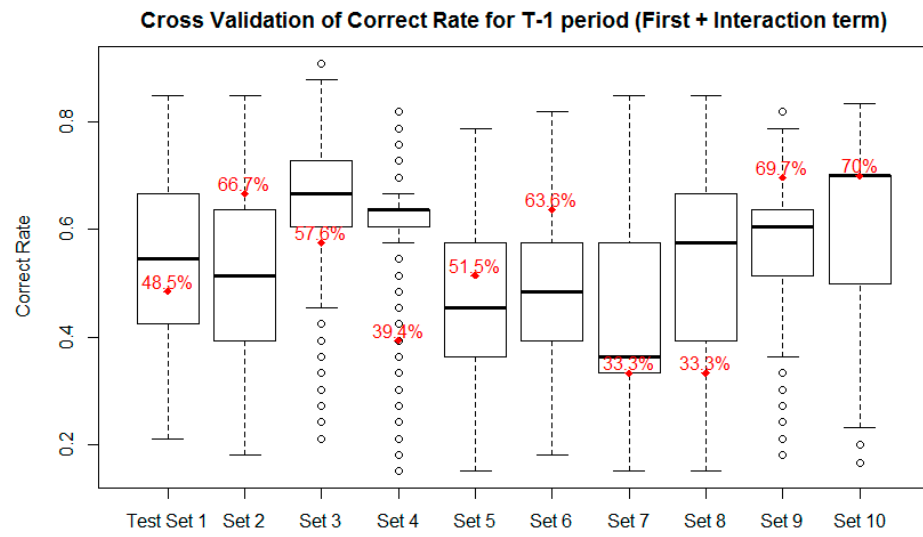


Figure 5. Cross-Validation of Correct Rate for T-1 period (First + Interaction term).

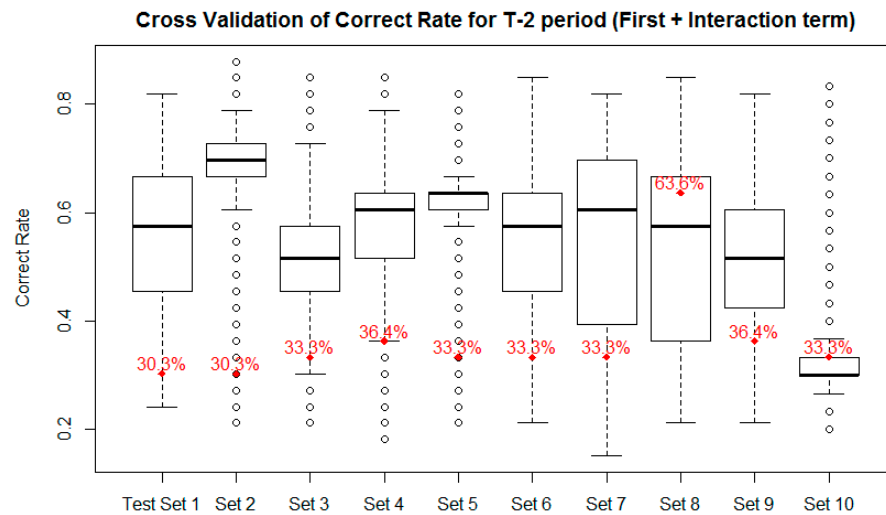


Figure 6. Cross-Validation of Correct Rate for T-2 period (First + Interaction term).

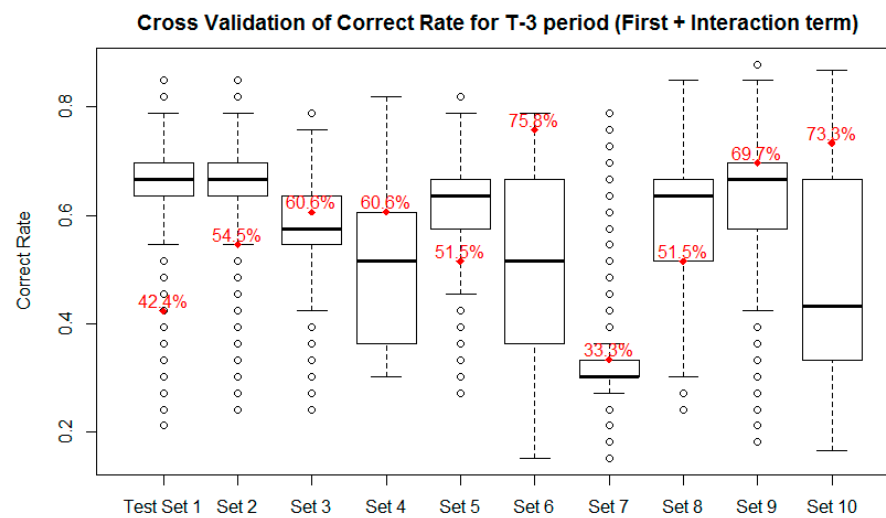


Figure 7. Cross-Validation of Correct Rate for T-3 period (First + Interaction term).

The results of the comparison are shown in Table 5. The T-test confirmed that there was a significant difference between the two, and the T-2 and T-3 phases were shown to have a higher accuracy rate based on the overall average, as seen at the end of Table 5.

Table 5. Model Comparison.

| Test Set | Period | Summary | First Order Correct Rate (Mean) | Interaction Term Correct Rate (Mean) | T-Test |
|----------|--------|---------|------------------------------------|---|-------------|
| Set 1 | T-1 | | 45.5% | 53.3% | −146.1 *** |
| | T-2 | | 48.6% | 54.9% | −121.4 *** |
| | T-3 | | 66.8% | 63.6% | 84.1 *** |
| Set 2 | T-1 | | 51.4% | 52.1% | −11.5 *** |
| | T-2 | | 50.2% | 68.4% | −372.6 *** |
| | T-3 | | 60.7% | 64.8% | −101.1 *** |
| Set 3 | T-1 | | 54.8% | 64.1% | −173.4 *** |
| | T-2 | | 47.7% | 51.5% | −89.0 *** |
| | T-3 | | 59.8% | 58.0% | 55.3 *** |
| Set 4 | T-1 | | 65.2% | 61.0% | 160.9 *** |
| | T-2 | | 55.7% | 56.5% | −17.5 *** |
| | T-3 | | 43.2% | 49.2% | −118.8 *** |
| Set 5 | T-1 | | 40.1% | 46.1% | −101.3 *** |
| | T-2 | | 34.6% | 62.4% | −1068.5 *** |
| | T-3 | | 51.4% | 61.8% | −252.0 *** |
| Set 6 | T-1 | | 60.4% | 48.4% | 236.2 *** |
| | T-2 | | 66.6% | 55.0% | 235.8 *** |
| | T-3 | | 43.7% | 50.6% | −113.5 *** |
| Set 7 | T-1 | | 50.5% | 43.4% | 107.4 *** |
| | T-2 | | 48.5% | 55.2% | −90.0 *** |
| | T-3 | | 66.3% | 33.1% | 928.0 *** |
| Set 8 | T-1 | | 34.1% | 54.6% | −361.2 *** |
| | T-2 | | 37.9% | 53.4% | −246.4 *** |
| | T-3 | | 69.4% | 59.2% | 253.0 *** |
| Set 9 | T-1 | | 54.2% | 55.8% | −31.8 *** |
| | T-2 | | 36.6% | 52.8% | −379.4 *** |
| | T-3 | | 59.8% | 61.1% | −22.3 *** |
| Set 10 | T-1 | | 42.2% | 60.6% | −269.5 *** |
| | T-2 | | 38.4% | 33.2% | 103.9 *** |
| | T-3 | | 69.7% | 47.7% | 329.4 *** |
| Total | T-1 | | 49.8% | 53.9% | −191.4 *** |
| | T-2 | | 46.5% | 54.3% | −369.4 *** |
| | T-3 | | 59.1% | 54.9% | 199.1 *** |

*** $p < 0.001$.

Comparisons of the predictive results from the traditional logistic and MCMC models regarding the 109 fraudulent companies are shown in Table 6. A logistic prediction of “1” indicates that fraud had occurred while “0” indicates no fraud. Using the MCMC method, there were 80,000 iterations for each sample, and the ratios in the fields represent the ratios of the 80,000 iterations predicted to be a fraud. According to Table 6, the MCMC provided clearly more information than the standard logistic model. For example, the 7th, 58th, 139th, 169th, and 322nd of the logistic model during the T-1 period was predicted to be normal; however, the MCMC’s predictions revealed fraud with over 76%, as highlighted in grey. Furthermore, the difference between the MCMC and the logistic model also occurs in the T-2 period in the 64th, 238th, 250th, and 256th samples. During the T-3 period, eight samples are predicted as normal, but the MCMC indicates it as fraud—such as the 202nd sample with 82.9%, or the 322nd sample with 88%.

Table 6. Fraud sample prediction comparison results.

| Test Set | Sample | T-1 | | T-2 | | T-3 | |
|----------|--------|----------|-------|----------|-------|----------|-------|
| | | Logistic | MCMC | Logistic | MCMC | Logistic | MCMC |
| Set 1 | 1 | 1 | 78.6% | 1 | 53.7% | 0 | 18.0% |
| | 4 | 1 | 42.1% | 1 | 4.9% | 0 | 5.8% |
| | 7 | 0 | 80.2% | 0 | 15.9% | 0 | 54.4% |
| | 10 | 0 | 29.8% | 1 | 99.2% | 1 | 7.0% |
| | 13 | 0 | 49.5% | 1 | 38.1% | 1 | 8.6% |
| | 16 | 0 | 30.4% | 1 | 60.1% | 1 | 14.7% |
| | 19 | 1 | 76.5% | 1 | 40.2% | 1 | 11.7% |
| | 22 | 0 | 55.7% | 1 | 43.1% | 1 | 23.4% |
| | 25 | 0 | 38.5% | 1 | 73.5% | 1 | 21.9% |
| | 28 | 0 | 24.3% | 1 | 71.5% | 1 | 32.8% |
| | 31 | 1 | 96.7% | 1 | 62.9% | 1 | 32.5% |
| Set 2 | 34 | 0 | 58.8% | 1 | 61.2% | 0 | 23.5% |
| | 37 | 0 | 37.0% | 1 | 87.2% | 0 | 77.1% |
| | 40 | 1 | 78.0% | 1 | 33.3% | 0 | 7.2% |
| | 43 | 1 | 56.0% | 1 | 24.3% | 1 | 5.7% |
| | 46 | 1 | 66.8% | 1 | 15.7% | 0 | 0.6% |
| | 49 | 1 | 43.0% | 1 | 23.3% | 0 | 16.6% |
| | 52 | 1 | 42.7% | 1 | 10.7% | 0 | 4.5% |
| | 55 | 1 | 39.4% | 1 | 35.0% | 0 | 11.5% |
| | 58 | 0 | 90.3% | 1 | 46.2% | 0 | 17.8% |
| | 61 | 0 | 59.4% | 1 | 9.0% | 0 | 82.3% |
| | 64 | 1 | 16.1% | 0 | 84.5% | 0 | 54.7% |
| Set 3 | 67 | 0 | 15.1% | 1 | 40.7% | 0 | 13.6% |
| | 70 | 0 | 22.8% | 1 | 82.2% | 0 | 10.3% |
| | 73 | 1 | 62.8% | 1 | 98.8% | 0 | 24.6% |
| | 76 | 0 | 4.6% | 1 | 16.2% | 0 | 25.9% |
| | 79 | 1 | 85.9% | 1 | 21.7% | 0 | 1.7% |
| | 82 | 0 | 25.4% | 1 | 74.7% | 0 | 28.7% |
| | 85 | 0 | 21.1% | 1 | 61.6% | 0 | 79.8% |
| | 88 | 1 | 68.8% | 1 | 79.3% | 0 | 26.0% |
| | 91 | 0 | 23.7% | 1 | 63.1% | 0 | 28.8% |
| | 94 | 0 | 62.1% | 1 | 79.3% | 0 | 13.0% |
| | 97 | 0 | 42.2% | 1 | 66.1% | 0 | 18.8% |
| Set 4 | 100 | 1 | 15.8% | 1 | 44.1% | 0 | 71.4% |
| | 103 | 1 | 5.3% | 1 | 44.1% | 0 | 70.5% |
| | 106 | 1 | 5.2% | 1 | 41.3% | 0 | 69.3% |
| | 109 | 1 | 1.6% | 1 | 42.4% | 0 | 62.9% |
| | 112 | 1 | 4.4% | 1 | 29.7% | 0 | 63.3% |
| | 115 | 1 | 21.9% | 1 | 43.5% | 1 | 72.8% |
| | 118 | 1 | 48.4% | 1 | 85.4% | 0 | 62.9% |
| | 121 | 1 | 6.6% | 1 | 26.5% | 0 | 51.5% |
| | 124 | 1 | 4.1% | 1 | 26.7% | 0 | 36.3% |
| | 127 | 1 | 12.7% | 1 | 33.1% | 0 | 59.9% |
| | 130 | 1 | 10.9% | 1 | 45.0% | 0 | 62.5% |
| Set 5 | 133 | 1 | 69.4% | 1 | 6.7% | 0 | 24.8% |
| | 136 | 1 | 38.1% | 1 | 4.0% | 0 | 11.4% |
| | 139 | 0 | 76.3% | 1 | 1.7% | 0 | 13.8% |
| | 142 | 1 | 19.9% | 1 | 0.8% | 0 | 6.8% |
| | 145 | 1 | 81.2% | 1 | 19.4% | 0 | 5.8% |
| | 148 | 0 | 9.0% | 1 | 6.6% | 0 | 17.7% |
| | 151 | 1 | 69.4% | 1 | 4.9% | 1 | 46.2% |
| | 154 | 1 | 47.0% | 1 | 22.3% | 0 | 6.0% |
| | 157 | 0 | 20.6% | 1 | 0.1% | 0 | 8.1% |
| | 160 | 1 | 48.5% | 1 | 2.1% | 0 | 11.2% |
| | 163 | 1 | 56.0% | 1 | 28.5% | 0 | 71.1% |

Table 6. Cont.

| Test Set | Sample | T-1 | | T-2 | | T-3 | |
|----------|--------|----------|-------|----------|-------|----------|-------|
| | | Logistic | MCMC | Logistic | MCMC | Logistic | MCMC |
| Set 6 | 166 | 1 | 70.2% | 1 | 59.6% | 1 | 74.3% |
| | 169 | 0 | 78.2% | 1 | 41.0% | 0 | 42.0% |
| | 172 | 0 | 50.8% | 1 | 54.8% | 0 | 66.0% |
| | 175 | 0 | 40.1% | 1 | 17.9% | 0 | 48.2% |
| | 178 | 0 | 54.1% | 1 | 39.8% | 0 | 58.5% |
| | 181 | 0 | 16.3% | 1 | 75.4% | 1 | 34.3% |
| | 184 | 0 | 56.3% | 1 | 24.5% | 0 | 48.3% |
| | 187 | 0 | 17.0% | 1 | 89.3% | 0 | 58.1% |
| | 190 | 0 | 54.3% | 1 | 69.6% | 0 | 41.8% |
| | 193 | 0 | 38.9% | 1 | 38.1% | 1 | 39.2% |
| 196 | 0 | 65.9% | 1 | 34.7% | 0 | 46.6% | |
| Set 7 | 199 | 1 | 87.3% | 1 | 13.0% | 1 | 81.3% |
| | 202 | 0 | 29.7% | 0 | 13.5% | 0 | 82.9% |
| | 205 | 1 | 67.7% | 1 | 33.3% | 1 | 92.4% |
| | 208 | 1 | 43.7% | 1 | 15.6% | 1 | 75.5% |
| | 211 | 1 | 90.5% | 1 | 61.0% | 1 | 97.9% |
| | 214 | 1 | 60.9% | 1 | 4.1% | 0 | 2.7% |
| | 217 | 1 | 60.0% | 1 | 44.5% | 1 | 89.5% |
| | 220 | 1 | 75.4% | 1 | 62.3% | 1 | 95.5% |
| | 223 | 1 | 54.8% | 1 | 28.3% | 1 | 83.5% |
| | 226 | 1 | 92.3% | 1 | 43.2% | 1 | 91.9% |
| 229 | 1 | 79.7% | 1 | 97.0% | 1 | 99.6% | |
| Set 8 | 232 | 1 | 39.8% | 0 | 45.1% | 1 | 76.5% |
| | 235 | 1 | 51.9% | 0 | 16.3% | 0 | 38.3% |
| | 238 | 1 | 47.3% | 0 | 96.1% | 1 | 61.3% |
| | 241 | 1 | 50.6% | 0 | 19.4% | 0 | 32.0% |
| | 244 | 1 | 32.4% | 1 | 55.5% | 0 | 30.4% |
| | 247 | 1 | 72.7% | 1 | 68.5% | 0 | 22.2% |
| | 250 | 0 | 53.8% | 0 | 71.5% | 1 | 42.7% |
| | 253 | 1 | 32.1% | 0 | 30.2% | 0 | 16.5% |
| | 256 | 1 | 84.2% | 0 | 98.8% | 0 | 37.1% |
| | 259 | 1 | 33.9% | 0 | 19.0% | 0 | 16.3% |
| 262 | 1 | 29.0% | 0 | 9.8% | 0 | 10.3% | |
| Set 9 | 265 | 0 | 56.1% | 1 | 71.7% | 0 | 31.4% |
| | 268 | 0 | 19.9% | 1 | 32.4% | 0 | 46.0% |
| | 271 | 0 | 31.4% | 1 | 87.3% | 0 | 29.5% |
| | 274 | 0 | 19.4% | 1 | 8.7% | 0 | 59.1% |
| | 277 | 1 | 50.8% | 1 | 70.1% | 0 | 52.1% |
| | 280 | 1 | 32.8% | 1 | 33.7% | 0 | 33.4% |
| | 283 | 0 | 10.8% | 1 | 84.1% | 1 | 37.7% |
| | 286 | 1 | 51.5% | 1 | 61.4% | 0 | 33.8% |
| | 289 | 1 | 29.1% | 1 | 52.0% | 0 | 37.9% |
| | 292 | 0 | 20.5% | 1 | 35.0% | 0 | 21.4% |
| 295 | 1 | 16.7% | 1 | 77.7% | 0 | 42.2% | |
| Set 10 | 298 | 0 | 21.5% | 1 | 98.7% | 1 | 83.4% |
| | 301 | 0 | 37.7% | 1 | 96.9% | 0 | 67.4% |
| | 304 | 0 | 35.2% | 1 | 72.2% | 0 | 51.5% |
| | 307 | 1 | 51.6% | 1 | 38.0% | 0 | 45.8% |
| | 310 | 0 | 27.1% | 1 | 96.8% | 0 | 43.6% |
| | 313 | 0 | 22.1% | 1 | 97.4% | 0 | 66.1% |
| | 316 | 0 | 17.8% | 0 | 31.9% | 1 | 3.4% |
| | 319 | 0 | 25.7% | 1 | 97.7% | 0 | 48.1% |
| | 322 | 0 | 88.8% | 1 | 99.4% | 0 | 88.0% |
| | 325 | 0 | 35.7% | 1 | 95.3% | 0 | 56.2% |

4.6. Model Error Analysis

Concerning the limitations of the models, the percentage of errors in the prediction results can be divided into false negatives and false positives. The error analysis results within the interaction models are shown in Figures 8–13. Besides, the box plots are shown in black, correspond to the sets of errors from the 80,000 iterations via the MCMC method and the red solid dots represent errors from the logistic model. In this study, I defined a false positive error as when a company was falsely accused of fraud. The false negatives occurred when a guilty company was judged to be innocent of fraud.

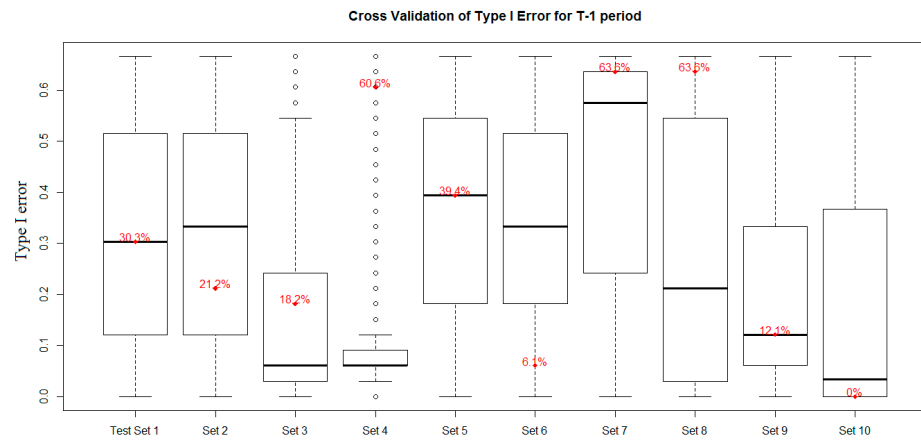


Figure 8. Cross-Validation of Type I Error for T-1 period.

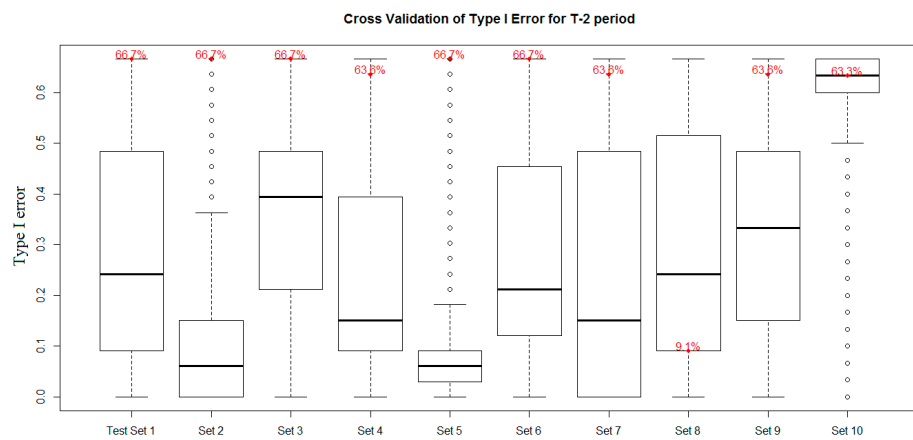


Figure 9. Cross-Validation of False Positive Errors in the T-2 Period.

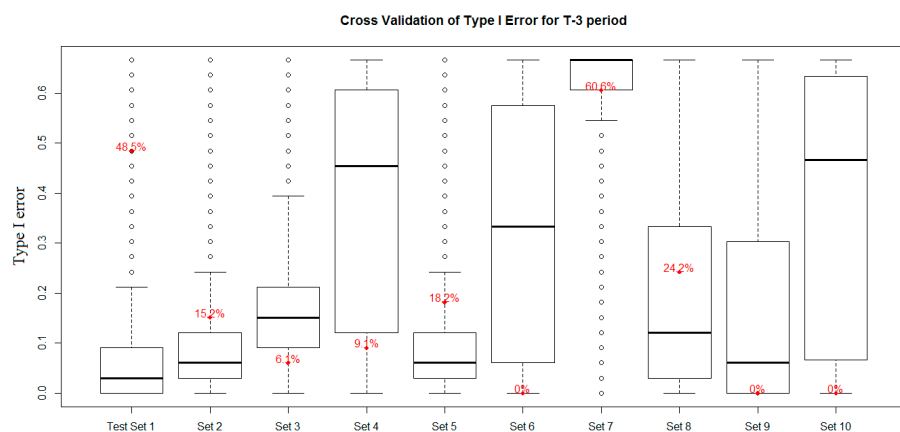


Figure 10. Cross-Validation of False Positive Errors in the T-3 Period.

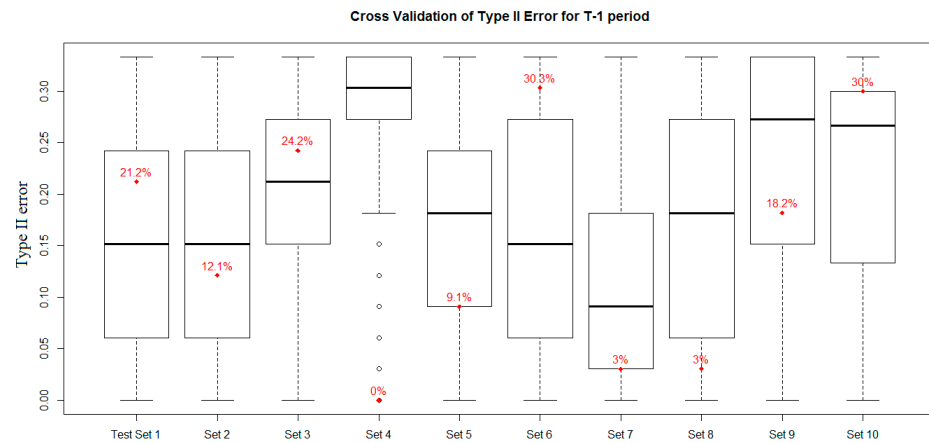


Figure 11. Cross-Validation of the False Negative Errors in the T-1 Period.

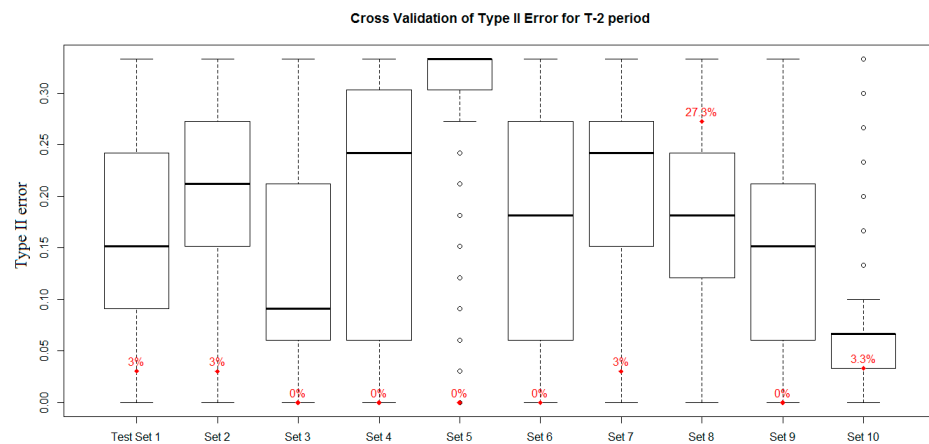


Figure 12. Cross-Validation of the False Negative Errors in the T-2 Period.

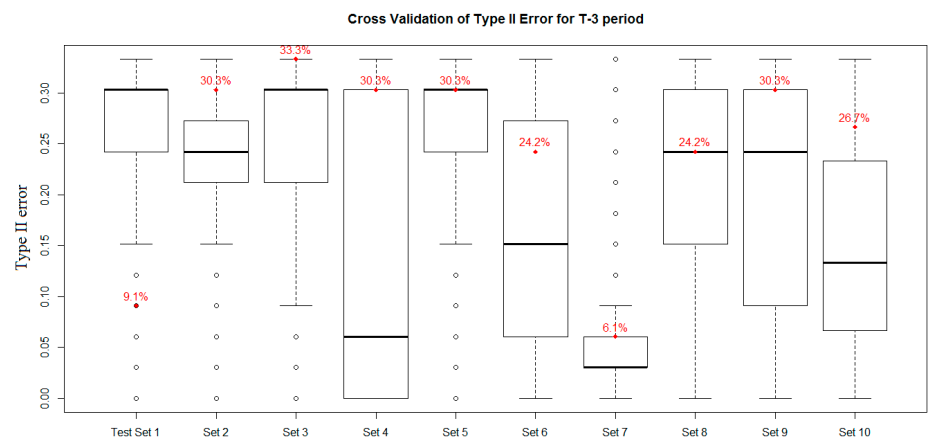


Figure 13. Cross-Validation of the False Negative Errors in the T-3 Period.

Furthermore, Figures 8–10 show the false positive errors in T-1 to T-3 periods. More than half of the 30 results (26 groups) using the logistic model deviated from the overall distribution of the 80,000 iterations. Figures 11–13 indicated false negatives in T-1, T-2, and T-3 periods, and 28 groups of these results from the logistic model deviated from the overall distribution.

The figures above clearly show that the results of the standard logistic prediction model also indicate an unstable state (i.e., overestimation or underestimation) regarding error analysis, meaning that if only the logistic model’s error results were analyzed, it would most likely result in a miscalculation of the error rate. Both the standard logistic and

the Bayesian probit model have their strengths and weaknesses. For example, the former may be too simplistic to handle such complicated data. Despite the complex nature of the Bayesian probit model, it could be used to correct the parametric estimation errors and reduce the problem of over- or under-confidence. As always, the best analysis method will depend on the problem that must be solved. Although the Bayesian probit model often yields clearer results, it is very complicated and expensive. Therefore, we recommend the integrated use of these two methods. The standard logistic model can be utilized for a preliminary analysis of the sample.

The Bayesian probit model will then be used for more precise calculations. Since over-fitting will interfere with the accuracy of the predictions yielded from the standard logistic model, it would not be as useful for real-world scenarios. However, as previously stated, the other model yields more accurate predictive results when the specific fitting of the correct model and data are used. These elements will be processed through the Bayesian probit model to take advantage of its more realistic predictive power, and also provide a visual component to help users better understand the distribution of prediction values. Above all, if the logistic model is used for prediction, a single result represents only one prediction point within the distribution space. However, if the MCMC model is used, multiple iterations may be used to offset the uncertainty of its parameters (dispersion of the predicted result). Thus, the MCMC model may be more appropriate for helping researchers understand the complexity of corporate fraud.

5. Conclusions and Recommendations

5.1. Theoretical Implications

In this study, we primarily employed the standard logistic model supplemented by Bayesian inference to counteract the uncertainty of model parameters. This study may be the first to use the boxplot to visualize the effects of model uncertainty and help users to make decisions based on the simulation results of model coefficients. Based on the proposed method, we also can eliminate the bias of parametric estimation for regular statistical models. In fact, the standard logistic model better revealed the analytical results while the Bayesian probit model with parameters via the MCMC showed a stable convergence. We also found that, unlike the standard logistic model, the distribution of unknown variables cannot be expressed in closed-form, and must be referred to as a simulated sample to accurately interpret the exact distribution value of the parameters.

Combining these two models to analyze the data yielded ideal predictive results. We found that the predictive power of the standard logistic model was stronger than that of the Bayesian probit model, which was more appropriate for approximating the maximum value. But the predictive power of the standard logistic model is unstable because the parametric estimation bias is inherent within the model. In this study, we used the MCMC model to calculate an unbiased estimation to enhance the predictive power and ameliorate the effects of model uncertainty.

5.2. Implications for Managers

For the investigation of fraud, the predictive results from the standard logistic model tended to be overly optimistic. However, the Bayesian probit model will significantly drive up the cost of analysis. Thus, although the full-range application of this model is ideal, it is not practical in the real world. For this reason, we suggest the integrated use of both models for the detection of fraud. In this way, the weakness of over-fitting would balance the unfitted model and data. After preliminary sorting of the data, the Bayesian probit model could be used for more precise calculations and would provide not only the prediction value of the responses but also possible ranges of these responses via a simple plot. This can help users to make informed decisions. In this way, the strengths of both models can be retained and utilized to their best advantage. This system would be much more accurate than applying the logistic model on its own to predict corporate fraud.

In this study, both models were run independently. The findings from both models using the same set of data unanimously indicated that the data from two years before the fraud occurred could most effectively predict this crime. As such, we can infer that indirect signs of fraud would begin to surface two years before it would become obvious. Therefore, issues related to corporate fraud, particularly fraudulent financial statements, not only require impeccable professional ethics and patience to correct the problem, but also a viable model that will allow for systematic analysis and reduce false accusations of fraud. Accordingly, companies that have been wrongly accused could be freed from unnecessary legal trouble, and these resources could be used more efficiently elsewhere. Most importantly, it could accurately detect companies that are engaged in acts of fraud. This would also help to protect the rights and privileges of the stakeholders and maintain stability within the market. Besides, the limitations of the proposed method still exist, such as the cost of analysis due to computationally expensive posterior distributions in the MCMC. In addition, the proposed model can be applied to the multinomial probit model in future studies. Further studies can be explored using other techniques to increase the efficiency of the MCMC algorithm, such as [56,57].

Author Contributions: Conceptualization, S.-H.T. and T.S.N.; formal analysis, S.-H.T. and T.S.N.; investigation, S.-H.T. and T.S.N.; methodology, S.-H.T. and T.S.N.; supervision, S.-H.T.; validation, S.-H.T. and T.S.N.; visualization, S.-H.T. and T.S.N.; writing—original draft, S.-H.T.; writing—review and editing, S.-H.T. and T.S.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Corporate information data. Available at Taiwan Economic Journal.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Persons, O.S. Using financial statement data to identify factors associated with fraudulent financial reporting. *J. Appl. Bus. Res.* **1995**, *11*, 38–46. [CrossRef]
- Kaminski, K.A.; Wetzel, T.S.; Guan, L. Can financial ratios detect fraudulent financial reporting? *Manag. Audit. J.* **2004**, *19*, 15–28. [CrossRef]
- ACFE. Report to the Nations on Occupational Fraud and Abuse. Available online: <https://www.acfe.com/report-to-the-nations/2020/> (accessed on 20 June 2021).
- Bologna, J.; Lindquist, R.J.; Wells, J.T. *The Accountant's Handbook of Fraud and Commercial Crime*; Wiley: New York, NY, USA, 1993.
- Mitnick, B.M. The theory of agency. *Public Choice* **1975**, *24*, 27–42. [CrossRef]
- Song, J.; Wang, R.; Cavusgil, S.T. State ownership and market orientation in China's public firms: An agency theory perspective. *Int. Bus. Rev.* **2015**, *24*, 690–699. [CrossRef]
- Schipper, K. Earnings management. *Account. Horiz.* **1989**, *3*, 91.
- Healy, P.M.; Wahlen, J.M. A review of the earnings management literature and its implications for standard setting. *Account. Horiz.* **1999**, *13*, 365–383. [CrossRef]
- Cressey, D.R. *Other People's Money: A Study of the Social Psychology of Embezzlement*; Free Press: Glencoe, IL, USA, 1953.
- Vousinas, G.L. Advancing theory of fraud: The SCORE model. *J. Financ. Crime* **2019**, *26*, 372–381. [CrossRef]
- Brennan, N.M.; McGrath, M. Financial statement fraud: Some lessons from US and European case studies. *Aust. Account. Rev.* **2007**, *17*, 49–61. [CrossRef]
- Skousen, C.J.; Smith, K.R.; Wright, C.J. Detecting and predicting financial statement fraud: The effectiveness of the fraud triangle and SAS No. 99. In *Corporate Governance and Firm Performance*; Hirschey, M., John, K., Makhija, A.K., Eds.; Emerald Group Publishing Limited: Bingley, UK, 2009; pp. 53–81.
- Hollow, M. Money, morals and motives: An exploratory study into why bank managers and employees commit fraud at work. *J. Financ. Crime* **2014**, *21*, 174–190. [CrossRef]
- Altman, E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **1968**, *23*, 589–609. [CrossRef]
- Beasley, M.S. An empirical analysis of the relation between the board of director composition and financial statement fraud. *Account. Rev.* **1996**, *71*, 443–465.
- Persons, O.S. The relation between the new corporate governance rules and the likelihood of financial statement fraud. *Rev. Account. Financ.* **2005**, *4*, 125–148. [CrossRef]

17. Tan, D.T.; Chapple, L.; Walsh, K.D. Corporate fraud culture: Re-examining the corporate governance and performance relation. *Account. Financ.* **2017**, *57*, 597–620. [[CrossRef](#)]
18. Xie, B.; Davidson, W.N., III; DaDalt, P.J. Earnings management and corporate governance: The role of the board and the audit committee. *J. Corp. Financ.* **2003**, *9*, 295–316. [[CrossRef](#)]
19. Francis, J.R. What do we know about audit quality? *Br. Account. Rev.* **2004**, *36*, 345–368. [[CrossRef](#)]
20. Hribar, P.; Kravet, T.; Wilson, R. A new measure of accounting quality. *Rev. Account. Stud.* **2014**, *19*, 506–538. [[CrossRef](#)]
21. Givoly, D.; Hayn, C. The changing time-series properties of earnings, cash flows and accruals: Has financial reporting become more conservative? *J. Account. Econ.* **2000**, *29*, 287–320. [[CrossRef](#)]
22. Kamarudin, K.A.; Ismail, W.A.W.; Mustapha, W.A.H.W. Aggressive financial reporting and corporate fraud. *Procedia Soc. Behav. Sci.* **2012**, *65*, 638–643. [[CrossRef](#)]
23. Lin, Y.-J. A Study of the Corporate Fraud Early Warning Models. Master's Thesis, National Chung Hsing University, Taichung, Taiwan, 2014.
24. Leith, C. Theoretical skill of Monte Carlo forecasts. *Mon. Weather Rev.* **1974**, *102*, 409–418. [[CrossRef](#)]
25. Amerstorfer, T.; Hinterreiter, J.; Reiss, M.A.; Möstl, C.; Davies, J.A.; Bailey, R.L.; Weiss, A.J.; Dumbović, M.; Bauer, M.; Amerstorfer, U.V.; et al. Evaluation of CME Arrival Prediction Using Ensemble Modeling Based on Heliospheric Imaging Observations. *Space Weather* **2021**, *19*, e2020SW002553. [[CrossRef](#)]
26. Buonaguidi, B.; Mira, A.; Bucheli, H.; Vitanis, V. Bayesian Quickest Detection of Credit Card Fraud. *Bayesian Anal.* **2021**, *1*, 1–30. [[CrossRef](#)]
27. Tseng, S.-H.; Kang, H.-Y.; Chen, H.-Y. A Test-Bed to Compare Alternative Bayesian Regression Formulations And An Application Of Cnc Milling Roughness Minimization. *J. Qual.* **2018**, *25*, 241–257.
28. Perols, J. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Audit. J. Pract. Theory* **2011**, *30*, 19–50. [[CrossRef](#)]
29. Fanning, K.M.; Cogger, K.O. Neural network detection of management fraud using published financial data. *Intell. Syst. Account. Financ. Manag.* **1998**, *7*, 21–41. [[CrossRef](#)]
30. Kirkos, E.; Spathis, C.; Manolopoulos, Y. Data mining techniques for the detection of fraudulent financial statements. *Expert Syst. Appl.* **2007**, *32*, 995–1003. [[CrossRef](#)]
31. Dong, W.; Liao, S.; Zhang, Z. Leveraging financial social media data for corporate fraud detection. *J. Manag. Inf. Syst.* **2018**, *35*, 461–487. [[CrossRef](#)]
32. Liu, C.; Chan, Y.; Kazmi, S.H.A.; Fu, H. Financial fraud detection model: Based on random forest. *Int. J. Econ. Financ.* **2015**, *7*, [[CrossRef](#)]
33. Baesens, B.; Höppner, S.; Verdonck, T. Data engineering for fraud detection. *Decis. Support Syst.* **2021**, 113492, in press.
34. Altman, E.I.; Iwanicz-Drozdowska, M.; Laitinen, E.K.; Suvas, A. Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model. *J. Int. Financ. Manag. Account.* **2017**, *28*, 131–171. [[CrossRef](#)]
35. Summers, S.L.; Sweeney, J.T. Fraudulently misstated financial statements and insider trading: An empirical analysis. *Account. Rev.* **1998**, *73*, 131–146.
36. Imhoff, G. Accounting quality, auditing and corporate governance. *Audit. Corp. Gov.* **2003**. [[CrossRef](#)]
37. Desai, M.A. The degradation of reported corporate profits. *J. Econ. Perspect.* **2005**, *19*, 171–192. [[CrossRef](#)]
38. Davidson, R.; Goodwin-Stewart, J.; Kent, P. Internal governance structures and earnings management. *Account. Financ.* **2005**, *45*, 241–267. [[CrossRef](#)]
39. Perols, J.L.; Lougee, B.A. The relation between earnings management and financial statement fraud. *Adv. Account.* **2011**, *27*, 39–53. [[CrossRef](#)]
40. Lennox, C.; Pittman, J.A. Big Five audits and accounting fraud. *Contemp. Account. Res.* **2010**, *27*, 209–247. [[CrossRef](#)]
41. Nusinovi, S.; Tham, Y.C.; Yan, M.Y.C.; Ting, D.S.W.; Li, J.; Sabanayagam, C.; Wong, T.Y.; Cheng, C.-Y. Logistic regression was as good as machine learning for predicting major chronic diseases. *J. Clin. Epidemiol.* **2020**, *122*, 56–69. [[CrossRef](#)] [[PubMed](#)]
42. O'Brien, S.M.; Dunson, D.B. Bayesian multivariate logistic regression. *Biometrics* **2004**, *60*, 739–746. [[CrossRef](#)] [[PubMed](#)]
43. Polson, N.G.; Scott, J.G.; Windle, J. Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Am. Stat. Assoc.* **2013**, *108*, 1339–1349. [[CrossRef](#)]
44. Sanchez-Lengeling, B.; Roch, L.M.; Perea, J.D.; Langner, S.; Brabec, C.J.; Aspuru-Guzik, A. A Bayesian approach to predict solubility parameters. *Adv. Theory Simul.* **2019**, *2*, 1800069. [[CrossRef](#)]
45. Ghosh, J.; Li, Y.; Mitra, R. On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Anal.* **2018**, *13*, 359–383. [[CrossRef](#)]
46. Gerlach, R.; Bird, R.; Hall, A.D. *A Bayesian Approach to Variable Selection in Logistic Regression with Application to Predicting Earnings Direction from Accounting Information*; School of Finance and Economics, University of Technology Sydney: Sydney, Australia, 2000.
47. Jackman, S. *Bayesian Analysis for the Social Sciences*; John Wiley & Sons: New York, NY, USA, 2009; Volume 846.
48. Rossi, P.E.; Allenby, G.M.; McCulloch, R. *Bayesian Statistics and Marketing*; John Wiley & Sons: New York, NY, USA, 2012.
49. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; John Wiley & Sons: New York, NY, USA, 2013; Volume 398.

50. Peduzzi, P.; Concato, J.; Kemper, E.; Holford, T.R.; Feinstein, A.R. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* **1996**, *49*, 1373–1379. [[CrossRef](#)]
51. Gilks, W.R.; Richardson, S.; Spiegelhalter, D. *Markov Chain Monte Carlo in Practice*; Chapman and Hall/CRC: Boca Raton, FL, USA, 1995.
52. Geman, S.; Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *Readings in Computer Vision*; Elsevier: Amsterdam, The Netherlands, 1987; pp. 564–584.
53. Von Toussaint, U. Bayesian inference in physics. *Rev. Mod. Phys.* **2011**, *83*, 943. [[CrossRef](#)]
54. Bernstein, L.A. *Analysis of Financial Statements*; Irwin Professional Publishing: Chicago, IL, USA, 1993.
55. Allen, T.T.; Tseng, S.H. Variance plus bias optimal response surface designs with qualitative factors applied to stem choice modeling. *Qual. Reliab. Eng. Int.* **2011**, *27*, 1199–1210. [[CrossRef](#)]
56. Joseph, V.R.; Wang, D.; Gu, L.; Lyu, S.; Tuo, R. Deterministic sampling of expensive posteriors using minimum energy designs. *Technometrics* **2019**, *61*, 297–308. [[CrossRef](#)]
57. Fielding, M.; Nott, D.J.; Liang, S.-Y. Efficient MCMC schemes for computationally expensive posterior distributions. *Technometrics* **2011**, *53*, 16–28. [[CrossRef](#)]