Profit-based Logistic Regression: A Case Study in Credit Card Fraud Detection

Azamat Kibekbaev, Ekrem Duman
Industrial Engineering Department
Özyeğin University
Istanbul, Turkey
E-mail: kibekbaev.azamat@ozu.edu.tr , ekrem.duman@ozyegin.edu.tr

Abstract— **Credit card fraud is a serious and growing problem which became increasingly rampant in recent years. In practice, many predictive models are used to identify fraudulent transactions. In this study, we developed a new profit-based logistic regression model. In order to do this, we modified the cost function in Maximum Likelihood Estimator (MLE) by changing its values according to the profit of each instance. We did this in four different scenarios and tested the results on real-life data of credit card transactions from an international Turkish bank. According to our findings, original Logistic Regression (LR) has the best performance in terms of TP rate. In terms of saving or net profit, profit-based LR scenarios outperformed others.**

Keywords-Fraud detection; Profit-based Logistic regression; MLE; cost function.

I. INTRODUCTION

Logistic Regression (LR) [17] is now widely used in credit scoring and credit card fraud more often than discriminant analysis because of the improvement of the statistical software for logistic regression. Moreover, LR is based on an estimation algorithm that requires less assumptions (assumption of normality, assumption of linearity, assumption of homogeneity of variance) than discriminant analysis. Prior work in related areas has estimated logit models (logit regression or logistic regression) of fraudulent claims in insurance, food stamp programs, and so forth [3][7][10]. It has been argued that identifying fraudulent claims is similar in nature to several other problems in real life including medical and epidemiological problems [13].

In credit card fraud detection, the dependent variable would take on a value of 0 (legitimate transaction) or 1 (fraudulent transaction). In this study, our dependent variable is binary and we estimate a LR model to predict fraud using primary and derived attributes as independent variables. In literature, a commonly used technique to detect credit fraud is LR. Such an econometric tool, together with the above mentioned techniques, is mostly employed within the credit scoring process to help institutions and organizations decide whether to issue credit to consumers who apply for it [1][4][5][6][16].

According to literature, Persons [12] developed a stepwise logistic regression model and provided evidence that accounting data is useful in detecting fraudulent financial reporting. Summer and Sweeney [15] report that a logistic model including insider trading variables differentiates between fraud and non-fraud firms. Lee, Ingram and Howard [9] document that a self-developed LR model has greater predictive ability when including the excess of cash flow over earnings as an explanatory variable, compared to only utilizing traditional financial statement variables. Bell and Carcello [2] construct a LR model based on multiple fraud-risk factors. They find that their relatively simple model consisting of several corporate governance and performance variables successfully differentiates between fraudulent and non-fraudulent observations. On the other hand, Kaminski et al. [8] present evidence that two regression models solely relying on basic financial ratios have limited use in detecting fraudulent financial statements. Sanjeev et al. [14] evaluated support vector machines and random forests, together with the LR, as part of an attempt to better detect credit card fraud. Random forests demonstrated overall better performance across performance measures.

In recent years, among all pattern recognition models, LR has become one of the outstanding linear algorithms with various applications from thrift failures and stock price predictions to bankruptcy prediction. Most of the previous studies have focused on cost of misclassification because in most of the problems, correct classification has no profit and there are just equal or different costs for different types of misclassifications. In above example regarding diagnosis problems, there are different costs for various misclassifications of healthy and unhealthy people. However, in most of the business problems, there is a cost-benefit wise perspective because correct classifications have some kinds of profit. For example, in "credit card fraud" if the base scenario is to take all of the instances as legitimate, if a model correctly detects a fraudulent transaction, it will save the accessible limit of the card and consequently will save it. In the direct marketing context, if a model correctly detects a potential customer for a campaign, there will be a profit of gaining that customer. Due to aforementioned reasons, in most of business problems, we have to develop a profit-cost wise prediction model. In the original version of LR, all of the misclassifications have same costs, which is not a realistic assumption in most of the real-world problems. For instance, in patient diagnosis problems,

ISBN: 978-1-61208-423-7

misclassification of an unhealthy as healthy is more risky and costly than misclassification of a healthy person as unhealthy. This issue motivated most of researchers to investigate the effect of different misclassification costs on classification models. For this reason, most of the works are related to cost-sensitive LR.

The remainder of the paper is organized as follows: the next section presents a brief literature survey on LR. Section 3 outlines modified error function or profit-based LR which takes the individual net profit into account and four applicable scenarios are presented to generate individual weights. Section 4 introduces the experimental results and discussions. Finally, Section 5 draws the conclusions of the study and indicates some possible future work areas.

II. Original and Profit-based Logistic Regression

LR is a statistical classification technique that has been developed in 1940’s and since then has been widely used in real life. It is similar to a linear regression model but is suited to models where the dependent variable is dichotomous. LR is often used when the dependent variable takes only two values and the independent variables are continuous, categorical, or both. The goal in LR is to find the best fitting, and most parsimonious model, to describe the relationship between a response or outcome variable, and a set of explanatory or predictor variables. LR model predicts the probability of occurrences, so if the odds of occurrences are higher than fifty percent, then the prediction will be assigned to class denoted by binary variable “1”, if less it is class “0”. The LR model is [18]:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \qquad (1)$$

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \qquad (2)$$

$$P(y = 1 \mid x; \theta) = h_\theta(x)$$
$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x) \qquad (3)$$

where the θ_i’s are the parameters and x_i are independent variables. Then, we can reformulate it as:

$$g(z) = \frac{1}{1 + e^{-z}} \qquad (4)$$

is called the logistic function or the sigmoid function as shown in Figure 1:

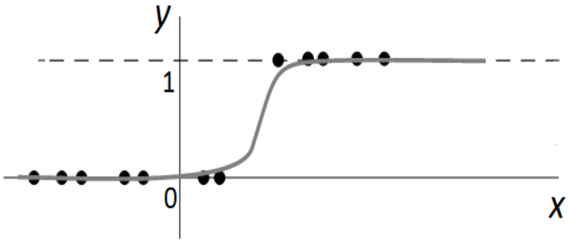

Figure 1. Sigmoid function

Then, we can write it more compactly as:

$$P(y \mid x; \theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y} \qquad (5)$$

Assuming that, the *m* training examples were generated independently, likelihood of the parameters will be:

$$\begin{aligned} L(\theta) &= p(\vec{y} \mid X; \theta) \\ &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) \\ &= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}} \end{aligned} \qquad (6)$$

It will be easier to maximize the log likelihood:

$$\begin{aligned} \ell(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)})) \end{aligned} \qquad (7)$$

After this, we now have to solve the maximization of likelihood. We used Newton’s method [19] (also called the Newton-Raphson method) given by:

$$\theta = \theta - H^{-1} \nabla_\theta l(\theta) \qquad (8)$$

where, $\nabla\theta\ell(\theta)$ is, as usual, the vector of partial derivatives of $\ell(\theta)$ with respect to the θ_i ’s; and H is an n-by-n matrix of second partial derivatives (actually, n +1-by-n + 1, assuming that we include the intercept term) called the Hessian:

$$H_{ij} = \frac{\partial^2 l(\theta)}{\partial\theta_i \partial\theta_j} \qquad (9)$$

Newton’s method typically enjoys faster convergence than (batch) gradient descent, and requires much less iteration to get very close to the minimum. The aim of Maximum Likelihood Estimator is to find the parameter values that make the observed data most likely to be predicted.

This paper proposes a new error function which modifies the original cost function to increase the total net profit. In this study, we defined four different scenarios to modify the error function and focused on profitability in the model building step. The key contribution entails that the proposed framework incorporates individual costs and benefits relevant for a business setting, as opposed to the current practice, which focuses on the statistical properties of classification algorithm. It seems obvious that these benefits and losses originating from correct and incorrect classifications should be taken into account. Note that allowing models to optimize the profitability criterion during the model construction step, leads to models with a higher performance in terms of profit although, it may decrease statistical performance of the model in comparison to previous models. Next section will explain our new modified error functions.

III. Profit-based LR Scenarios

Our main goal is to correctly classify the profitable instances as much as possible so that there is less decrease in the accuracy of detecting other instances (i.e. not profitable ones). For this reason, an indicator has been used in the error function to make the algorithm more sensitive to high profitable instances without affecting others. Accordingly, we used a multiplier to intensify the individual penalty of profitable false negatives (in CC Fraud, fraudulent misclassifications which their usable limit is more than average).

We can consider this modification from another point of view. A learning rate is user-defined value to determine how much the weights of examples can be modified at each iteration. We can assume that the learning rate has been modified to assign an appropriate individual penalty for each example and penalize the misclassified important examples considering their individual importance.

The indicator should indicate the profitable (important) instances using their attribute which shows the importance of instance which is Usable Limit (UL) in the context of credit card fraud and the customer revenue (balance) in direct marketing. Thus, indicator has been defined as:

$$P_i = \begin{cases} 1 & \mathit{if}\ UL_i > AvgUL \\ 0 & \mathit{otherwise} \end{cases} \qquad (10)$$

$$AvgUL = \frac{1}{n}\sum_{i=1}^{n} UL_i \qquad (11)$$

A. Scenario1

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log\left(h_\theta(x_i)\right) + (1-y_i) * \log\left(1-h_\theta(x_i)\right)\right) * \left(\frac{UL_i}{Avg\ UL}\right)^{P_i} \qquad (12)$$

where UL_i is the individual profit of instance *i* and $AvgUL$ is average usable limit of an instance. Our main goal is to correctly classify the profitable instances as much as possible with minimum decrease in the accuracy of detecting other instances.

B. Scenario2

As the ratio $\frac{UL_i}{Avg\ UL}$ in the previous scenario can give out large values it may cause instability in the model, so for the sake of making the multiplier not a very large value, we can use logarithm function in an alternative scenario. Hence, the penalty for each instance can be defined as:

$$P_k = \ln(1 + \frac{UL_i}{Avg\ UL}) \qquad (13)$$

The value of one inside the logarithm guarantees that the output will always be positive as the ratio $\frac{UL_i}{Avg\ UL}$ is a positive real number. The penalty function and weight updating equations can be expressed as:

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log\left(h_\theta(x_i)\right) + (1-y_i) * \log\left(1-h_\theta(x_i)\right)\right) * (\ln\left(1 + \frac{UL_i}{Avg\ UL}\right))^{P_i} \qquad (14)$$

C. Scenario3

This scenario is based on modified Fisher [11]. In this scenario, there is no indicator for profitable instances where all of the instances are given a weight related to their potential profit. The error function for this scenario is as follows:

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log\left(h_\theta(x_i)\right) + (1-y_i) * \log\left(1-h_\theta(x_i)\right)\right) * \left(1 + \frac{UL_i}{Avg\ UL}\right)^{1/2} \qquad (15)$$

D. Scenario 4

This scenario gives different weights for different instances considering their profit of correct classification. Instead of average usable limit we divided it by the maximum of limits. For this reason, this Max_LR error function is:

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log\left(h_\theta(x_i)\right) + (1-y_i) * \log\left(1-h_\theta(x_i)\right)\right) * \left(1+\frac{UL_i}{\max\{UL_i\}}\right) \quad (16)$$

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The credit card (CC) fraud data set has been gathered from a well-known Turkish bank and it contains 9243 transaction where 8304 of them are legitimate and 939 are fraudulent ones. In the empirical study of each data, the data set has been divided in a way that 2/3 proportion is used to train the model and 1/3 is used to test the trained model. Therefore, there are 313 fraudulent instances and 2817 legitimate ones in the test set. In all the scenarios, the train sets and test sets are the same. However, as the initial weights are generated randomly from standard normal distribution to cope for the effects of randomness related with the solution of train/test sets and the algorithm parameters. Also, each of the models has been run ten times and the average of runs is considered as classifiers' final performance.

In the context of credit card fraud, the most important profit-based attribute is the usable limit of each card. If we correctly detect fraudulent cases, we save their usable limit subject to a cost of contact. Let us consider the base scenario as the case where all transactions are supposed to be legitimate. It is a common approach for evaluating the profit of applying data mining algorithms. Then, the following expression demonstrates how to calculate the amount of net profit (saving) for each model:

$$NP = \sum_{i=1}^{N_{TP}}(UL_i - c) + \sum_{k=1}^{N_{FP}}(-c) \quad (17)$$

where c is the fixed cost for each alarm (cost of contacting the customer) and N_{TP} and N_{FP} indicate the number of true positives and false positives, respectively. As mentioned above, UL_i is the amount of profit gained when the instance i is classified correctly. The threshold has been changed from 0.5 to the number of cases (positives) in test set to show that in the top most probable instances, which of the classifiers is successful.

"*Saving*" measures the amount of profit in each model with threshold 0.5. The **"*Net profit in top n*"** (***n*** is the number of actual positives in test set) evaluates net profit when the cutoff point is output of top nth instance. This measure has an advantage that doesn't care about the number of total positives in each classifier, but it gives more importance to the actual number of positives detected in the first top positives in each model and sums their net profits.

Tables 1-3 illustrate the performances of the four scenarios and original LR on the given data set. According to statistical measure, original LR has the greatest TPR as it tries to correctly classify instances as much as possible where instance's profitability is not important. Also, profit-driven LR in 3rd scenario has also compatible TPR. However, in savings profit-based LR showed better performance (especially 3rd and 4th scenarios). In the average results, Modified Fisher scenario (3rd) has highest amount when threshold is on top 313th instance and Max_LR (4th) outperformed in total savings.

TABLE I. TRUE POSITIVE RATE

Scenario	TP rate		
	Min	Avg	Max
Original	**0,765**	**0,778**	**0,782**
1st	0,764	0,768	0,775
2nd	0,758	0,767	0,778
3rd	0,756	0,772	0,780
4th	0,763	0,769	0,774

TABLE II. TOTAL SAVINGS ON TEST SET

Scenario	Total Saving (%)		
	Min	Avg	Max
Original	0,730	0,762	0,798
1st	0,761	0,775	0,808
2nd	0,766	0,782	0,814
3rd	**0,780**	0,795	0,810
4th	0,770	**0,797**	**0,834**

TABLE III. TOP 10% SAVING ON TEST SET

Scenario	Saving (%) on top 313		
	Min	Avg	Max
Original	0,775	0,793	0,810
1st	0,775	0,800	0,827
2nd	0,787	0,804	0,820
3rd	**0,790**	**0,820**	0,840
4th	0,773	0,815	**0,846**

V. CONCLUSION AND FUTURE WORK

In this study, a novel profit-based logistic regression has been proposed which makes the classification considering all individual costs and profits of instances and consequently maximizes the total net profit captured from applying the classification model. For this purpose, we modified the logistic regression error function which is sensitive to instances' profitability's. Different scenarios have been proposed to generate weights (penalties) for modification of error function. All scenarios have been tested on a real-life fraud data set. In order to evaluate the classifiers, both TP rate and Savings performance metrics have been used. According to results, original LR has the best performance in terms of TP rate. While, in terms of saving profit-based LR (Modified Fisher and Max_LR) scenarios outperformed others.

As for the future research, we are working on models which assign an individual profit for the non-cases which have been classified correctly. As there is a variable cost of making a contact with each customer, they may get annoyed by this action of being contacted and there might be a cost of missing a customer and consequently missing his/her life time value or future profits.

ACKNOWLEDGMENT

With a deep sense of gratitude the authors would like to thank The Scientific and Technological Research Council of Turkey (TÜBİTAK) under Project No. 113M063.

REFERENCES

[1] H.A. Adbu, "An evaluation of alternative scoring models in private banking," Journal of Risk Finance, vol. 10 (1), 2009, pp. 38-53.

[2] T.B. Bell and J.V. Carcello, "Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting," Auditing: A Journal of Practice & Theory, vol. 19, 2000, pp. 169-184.

[3] C.R. Bollinger and M.H. David, "Modeling discrete choice with response error: food stamp participation," Journal of the American Statistical Association, vol. 92, 1997, pp. 827–835.

[4] J. Crook, and J. Banasik, "Does reject inference really improve the performance of application scoring models?" Journal of Banking & Finance, vol. 28 (4), 2004, pp. 857-874.

[5] V.C. Desai, J.N. Crook and J.G.A. Overstreet, "A comparison of neural networks and linear scoring models in the credit union environment," European Journal of Operational Research, vol. 95 (1), 1996, pp. 24-37.

[6] W.H. Greene, "Sample selection in credit-scoring models," Japan and the World Economy, vol. 10, 1998, pp. 299-316.

[7] J.A. Hausman, J. Abrevaya and F.M. Scott-Morton, "Misclassification of a dependent variable in a discrete-response setting," Journal of Econometrics, vol. 87, 1998, pp.239–269.

[8] K.A. Kaminski, T.S. Wetzel and L. Guan, "Can financial ratios detect fraudulent financial reporting?" Managerial Auditing Journal,vol. 19, 2004, pp. 15-28.

[9] T.A. Lee, R.W. Ingram and T.P. Howard, "The Difference between Earnings and Operating Cash Flow as an Indicator of Financial Reporting Fraud." Contemporary Accounting Research, vol. 16, 1999, pp. 749-786.

[10] M. Artis, M. Ayuso and M. Guillen, "Detection of automobile insurance fraud with discrete choice models and misclassified claims," The Journal of Risk and Insurance, vol. 69 (3), 2002, pp. 325–340.

[11] N. Mahmoudi and E. Duman, "Detecting credit card fraud by Modified Fisher Discriminant Analysis," Expert Syst. Appl., Nov. 2014.

[12] O.S. Persons, "Using financial statement data to identify factors associated with fraudulent financial reporting," Journal of Applied Business Research, vol. 11, 1995, pp. 38-46.

[13] S.B. Caudill, M. Ayuso and M. Guillen, "Fraud detection using a multinomial logit model with missing information," The Journal of Risk and Insurance, vol. 72 (4), 2005, pp. 539–550.

[14] J. Sanjeev, M. Guillen and J.C. Westland, "Employing transaction aggregation strategy to detect credit card fraud," Expert Systems with Applications, vol. 39, 2012, pp. 12650–12657.

[15] S.L. Summers and J.T. Sweeney, "Fraudulently misstated financial statements and insider trading: An empirical analysis," The Accounting Review,vol. 73, 1998, pp. 131-146.

[16] L.C.A. Thomas, "Survey of credit and behavioural scoring: forecasting financial risk of lending to consumers," International Journal of Forecasting, vol. 16 (2), 2000, pp. 149-172.

[17] C. Spathis, "Detecting False Financial Statement Using Published Data: Some Evidence from Greece," Managerial Auditing Journal, vol 17, April 2002, pp.179-191.

[18] D.W. Hosmer and S. Lemeshow, " Applied Logistic Regression (2nd ed.)," Wiley, 2000.

[19] P. Komarek and A. W. Moore, "Making logistic regression a core data mining tool: A practical investigation of accuracy, speed, and simplicity," Robotics Institute, Carnegie Mellon University, 2005.