Maximizing Profit Efficiency as Overall Evaluation Criterion for Fraud Control

Authors
Co-authored by Justin Wang and Jay Nanduri, with contributions from Justin Hobart and Harish Jayanti.

Introduction
eCommerce fraud costs retailers approximately $4 billion***** each year, according to the most recent results of an annual survey conducted by Cybersource, a provider of electronic payment and risk management services. Since eCommerce is a "card not present" scenario, merchants are responsible for fraudulent losses: merchants need to return the collected funds to the card issuing banks, which is known as a chargeback, when card holders report the transactions are fraudulent (unauthorized usage).

To control fraud costs, traditionally, financial instruments and credit card issuing banks use the chargeback rate as the measurement to evaluate the performance of Fraud Control. Since this metric penalizes missing fraud (false negatives) heavily, the strategies developed to improve chargeback rate tend to be over protective and only approve very low risk transactions.

As a result, many good transactions are rejected (false positives). Currently, in general, chargeback rates are lower than 1% while issuing banks reject higher than 15% of transactions. In the field of statistical classification in Machine Learning, more comprehensive measurements (e.g., accuracy, recall, or false positive rate) are introduced in a table of confusion (sometimes also called a confusion matrix).

Unfortunately, those measurements can be misleading when fraud attacks happen. They also do not take margin and cost of goods into the consideration, which are essential since the business goal is often to take the approach which can maximize net profit. We propose using Profit Efficiency (PE) as the standard measurement for Fraud Control.

There are three main advantages of adopting this measurement:

1. Maximizing profit efficiency leads to the strategies which yield maximal profit. For goods with higher cost and lower margin, the risk enforcement is more intensive and, on the other hand, for goods with lower cost and higher margin, we are more willing to take risk with a lighter risk enforcement.

2. Unlike other measurements which might be misleading when the business is under severe fraud attacks, profit efficiency honestly reflects the fact and shows the loss.

3. Optimizing profit efficiency is very straightforward. Comparing with balancing sensitivity/specificity or precision/recall, profit efficiency is easy to understand and implement.

Profit Efficiency Terminologies
In the business world, maximizing profit is usually the ultimate goal that management would like to reach. PE takes cost of the goods (COGS) into account and is calculated with the following logic:

- COGS = cost of goods including revenue sharing, loyalty, operations, and other costs
- Margin = Sale price - COGS
- PE = margin earned for an approved good transaction
- PE = -1*COGS for an approved bad transaction
- PE = 0 for a rejected transaction (here we ignore processing cost, since its negligible)

Confusion Matrix and Profit Calculation:

		Decision	
		Approve	**Reject**
Ground Truth	**Good**	True Negative	False Positive
		PE = margin	PE = 0
	Bad	False Negative	True Positive
		PE = - COGS	PE = 0

$$\text{Profit Efficiency} = \frac{\text{Profit Achieved}}{\text{Maximum Profit Achievable}} = \frac{\text{Margin(True Negatives)} - COGS(False\ Negatives)}{\text{Margin(True Negative)} + \text{Margin(False Positives)}}$$

Compare PE with other metrics

The performance indexes of a Risk system can have different values when the attack pattern changes. This section illustrates a numeric example to show how the commonly used metrics and PE reflect to a fraud attack. In a regular day, there are 110 fraudulent transactions attacking the platform, among which, 100 are caught and 10 of them can bypass the Risk checks. When a fraud attack happens, fraudsters stress the system by increasing attempts by 500% to improve the successes by 50%. This illustrates a typical fraud attack in the real world. Assuming the good traffic remains the same, in the following table, we can see False Positive Rate (FPR) does not change, Precision, Recall, and Accuracy have more favorite values while Chargeback Rate and PE ratio are less favorable.

Clearly Precision, Recall, and Accuracy are more sensitive in reflecting the number of fraudulent transactions caught (for the additional 505 attempts, 500 or 99.01% are caught comparing with the regular attack where 100 out of 110 or 90.91% are caught), while Chargeback Rate and PE ratio tend to reflect the fact that the number of approved bad transactions has increased from 10 to 15 which causes additional financial loss.

If we are to pick a monthly measurement for performance index and assume we have a "regular" month followed by an "under attack" month, it would be far from desirable to report a positive trend (which Precision, Recall, and Accuracy will do) when the financial loss does increase substantially. Focusing on the bright side that more fraudulent attempts are caught is not very convincing since it does not correctly reflect the business concerns. This is probably the main reason that most financial institutions and credit card issuing banks prefer to use Chargeback Rate as their tracking metric.

However, using Chargeback Rate as the key performance index leads to a very conservative strategy of approving transactions since, by definition, it almost does not penalize false positives in the field of Fraud Control where large majority of approved transactions are good. Today in eCommerce, issuing banks reject 15% of transactions and an internal study by Mastercard shows 80% of rejects are probably false positives. PE ratio provides a metric which takes both false positives and false negatives into account. It also successfully reports the trend which businesses desire to see, not even mentioning its major advantage of driving strategies which are willing to take more risk on low cost goods and a more conservative approach on high cost ones.

		A Regular Day	Under Attack
TN	Approved Good Txns	1000	1000
FP	Rejected Good Txns	5	5
TP	Rejected Bad Txns	100	600
FN	Approved Bad Txns	10	15
	Chargeback Rate	0.99%	1.48%
	FPR	0.50%	0.50%
	Precision	95.24%	99.17%
	Recall	90.91%	97.56%
	Accuracy	98.65%	98.77%
	PE Ratio (assuming 50% margin)	98.51%	98.01%

Other Applications
Like accuracy, which summarizes the four components of the confusion matrix into a single value, profit efficiency is also a metric which can evaluate a complex problem with a single number. In fact, profit efficiency can be viewed as an advanced version of weighted accuracy, which rewards/penalizes the correct/incorrect decision by the margin and cost of goods. This version of accuracy directly reflects business interests and can be intuitively understood. Note that, although true positives (TP) are not seen as part of the formula, they are implicitly included in the measurement.

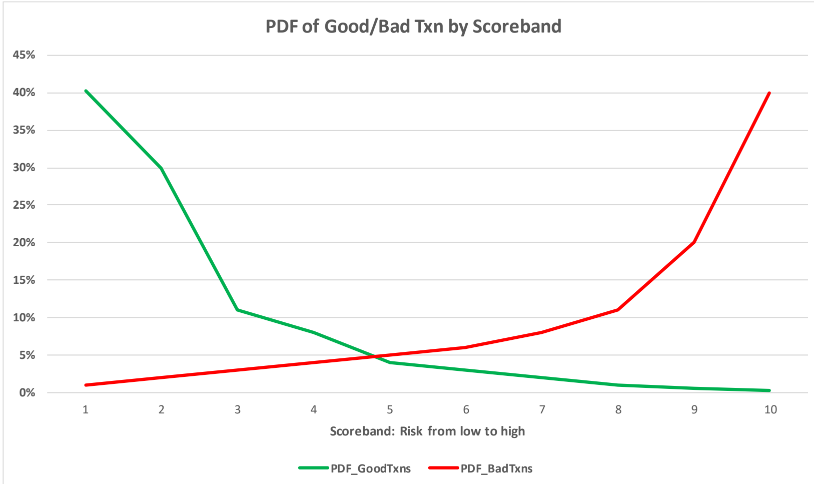
Although the concept can be easily applied to the other machine learning field and clinic trials, the challenge is how to quantify the loss or gain of false positives (FP), true negatives (TN), false negatives (FN), and true positives (TP). This is similar to restricting type I error and type II error for test hypothesis. People know Type I error is often less tolerant than Type II error; therefore, in practice, researchers usually allow 5% of Type I error and 20% of Type II error (or 80% power). However, 5% and 20% are picked arbitrarily without true business or physical meanings.

For the business or classification cases in which FP, TN, FN, and TP can be quantified, we encourage using the concept of (achieved)/(maximum achievable) as the overall evaluation criterion. This criterion indices the perfect execution with 100% and provides a measurement of how far the current execution is from perfect execution with a percentage.

Use Risk Score to Maximize Profit Efficiency
In eCommerce industry, the major online retailers, including Amazon, Apple, Google, and Microsoft, use machine learning models for fraud control. The common practice is training a supervised machine learning model based on historical data: rejecting the transactions with high predictive possibility of being a bad transaction and approve the others. By design, score is ranking ordered by fraudulent probability.

The figure below demonstrates the probability distribution function (PDF) of good transactions and bad transactions by score bands from low (lower chance of being fraudulent) to high (higher chance of being fraudulent). Transactions scored above the cutoff are rejected and below the cutoff are approved. Good transactions scored above the cutoff are false positives (incorrectly rejected), below the cutoff are true negatives (correctly approved); bad transactions scored above the cutoff are true positives (correctly rejected), and below the cutoff are false negatives (incorrectly approved).

Decision makers may choose the cutoff which maximizes the selected KPI. However, one fact often ignored by the decision makers is that, depending on the score distribution, the variance of the KPI value can be very different when different cutoff is chosen. A cutoff which optimizes KPI, if having large variance, has a high chance of introducing uncertainty and results in unexpected return when applied to real decisions.



***Source** = CSO Online, "E-commerce Fraud: The Latest Criminal Schemes": http://www.csoonline.com/article/2124192/malware-cybercrime/e-commerce-fraud--the-latest-criminal-schemes.html