1. Introduction

This chapter discusses how to assess the performance of a fraud detection system. Intuitively, the task seems simple. An ideal fraud detection system should maximize the number of correct classifications, and detect all fraudulent transactions. Hence, it is tempting to think that simply minimizing the proportion of misclassified transactions (a metric known as *mean misclassification error*) is the metric to optimize.

As we will show shortly, the mean misclassification error is a poor performance metric, due to the cost-sensitive and imbalanced nature of a fraud detection problem. A simple way to illustrate this is to observe that, for a transaction dataset with 0.1% fraudulent transactions, a dummy baseline model classifying all the transactions as genuine provide a very high accuracy of 0.99. This is widely acknowledged in the fraud detection literature, and other performance metrics are therefore commonly used [DPBC+17, Elk01, Tha20]. The most common ones are the *recall*, the *specificity*, the *precision*, the *F1 score*, the *AUC ROC*, and the *Average Precision*.

In the next sections, we will detail these metrics, and discuss their pros and cons. We will show that, despite their central role in assessing a fraud detection system, there is actually no consensus on what metric should be used.

The recall, specificity, precision, and F1 score metrics, also known as *threshold-based* metrics, have well-known limitations due to their dependence on a decision threshold which is difficult to determine in practice, and strongly depends on the business-specific constraints. They are often complemented with the AUC ROC, and more recently, the Average Precision (AP) metrics. The AUC ROC and AP metrics aim at assessing, with a single number, the performance for all possible decision thresholds, and are referred to as *threshold-free* metrics. The AUC ROC is currently the de-facto metric for assessing fraud detection accuracies [Cha09, DP15]. Recent research has however shown that this metric is also misleading for assessing highly imbalanced problems such as fraud detection [Mus19], and recommended using the Precision-Recall curve and AP metric instead [BEP13, SR15].

The chapter is structured as follows. Section 4.2 first presents fraud detection as a classification problem and details the main threshold-based metrics. Through a simple example, we show that the mean misclassification error is a bad indicator of performance, and motivate the use of alternative metrics such as recall, specificity, precision, and F1-score. Section 4.3 then discusses the use of threshold-free measures such as AUC ROC and AP, and show their benefits and limitations. Section 4.4 finally addresses the fraud detection problem from a more operational perspective and motivate the use of the *Card Precision top-k metric*.

Print to PDF

Previous
6. Summary

Next
2. Threshold-based metrics