Confidence Based Performance Estimation

Assessing the Performance of Machine Learning Models on Financial Data without Ground Truth

Applied Data Science Master Thesis

Utrecht University

Auditdienst Rijk
Ministerie van Financiën

Author:	Alex Essaijan
Student number:	7247044
Date of submission:	July 7, 2022
Version:	Final thesis
Utrecht University Supervisor:	Dr. Sjoerd Dirksen
Utrecht University Second Supervisor:	Prof. dr. ir. Kees Oosterlee
External Supervisor:	Fré Vink
External Company	Dutch Central Government Audit Service
Qualification:	MSc Applied Data Science
Faculty:	Natural Science, Dept. of Informatics

Acknowledgements

First and foremost, I do like to express my gratitude to my supervisors Sjoerd Dirksen and Kees Oosterlee, for their patience and support in the completion of this project. Furthermore, I would like to take this opportunity to sincerely thank Fré Vink as supervisor from The Dutch Central Government Audit Service, for his valuable advice and support during this thesis process. Especially in navigating the systems and structures of the Ministry of Finance. Finally, I want to thank my fellow students, friends, and parents, who have inspired me from the start and offered their continuous support.

Abstract

In the field of machine learning, the evaluation of models typically involves training them on a specific dataset and assessing their performance on a separate test set. However, assessing their performance in real-world environments can be challenging, especially when there is a shortage of labeled data. This study focuses on estimating the performance of machine learning classifiers in financial audits, specifically on unseen accounting data. By employing the Confidence Based Performance Estimation methodology, accurate estimation of performance metrics can be achieved, considering both predicted labels and probabilities. These estimates can be made under the assumption that there is no concept drift, the model is well calibrated, and it exhibits consistent performance across all classes. The findings of this study have practical implications for auditors, offering insights into the feasibility and usability of integrating machine learning models into audit procedures. This enables auditors to make informed decisions regarding the adoption of these models. Furthermore, this research contributes to the field by emphasizing the importance of considering class discrepancies and promoting a data-driven approach to improve sampling methods beyond traditional random sampling. In future research, it would be valuable to address challenges such as multiclass calibration, class imbalance, threshold selection methods, and real-time monitoring of model performance. These areas of investigation would enhance the robustness and applicability of machine learning models in production settings.

Table of Contents

Acknowledgements 1

Abstract 2

1. Introduction 4

2. Related work 6

2.1 Machine Learning in the Audit Domain 6

2.2 Machine learning performance evaluation on unseen data 7

2.3 Model Calibration 8

2.4 Drift 9

3. Dataset Description 10

4. Methodology 11

4.1 Model Description 11

4.2 Probability Calibration 11

4.3 Scoring rules 12

4.3.1 Expected Calibration Error 12

4.3.2 Log Loss 13

4.3.3 Brier score 13

4.4 Performance Estimation 14

5. Results 15

6. Discussion and Conclusion 19

7. Limitation & future research 20

8. Bibliography 21

1. Introduction

In machine learning, the process of creating models typically involves dividing the data into a training and test set. The model is then trained on the training set and its performance evaluated on the test set. It is commonly assumed that these sets are representative of the production dataset, meaning that they have similar statistical properties and characteristics (Arbet et al., 2020). This assumption allows for the performance of the model on the test set to be used as a proxy for the model's performance on the production data. Since it allows for a quick and efficient evaluation and helps to save time and resources by not having to evaluate the model every time it is used (Cinà et al., 2023).

While the assumption is that the training and test sets are representative of the production dataset is useful in evaluating the model's performance, it is not always accurate. Real-world environments may present different data distributions that the model has not been exposed to during training. As a result, the model's performance may vary significantly on new and unseen data, which may have different characteristics and patterns. A possible approach to mitigate this problem is to gather new labeled test data to validate the model (Vartak, 2021). However, obtaining new labeled data each time a model is used in production, may be prohibitively expensive. But without a reliable estimate of the model's performance on such data, it can be challenging to make informed decisions or trust the model's results. Thus, the ability to accurately predict changes in a model's performance using only unlabeled data becomes critical in real-world applications, certainly when the availability of labeled data is limited.

One of the places where the availability of labeled data is limited and costly to acquire, is financial audits. This presents a unique challenge when it comes to developing reliable machine learning models. In the context of audits, there is typically no labeled data available in advance, as the transactions to be audited still need to be checked. This makes it difficult to validate the performance of machine learning models on the same data that they will be audited, and the representative data assumption might not hold. As a result, there is uncertainty in the model's performance, making it unreliable to use since it potentially leads to financial misstatements. This explains why the majority of external audits are still sample-driven (Appelbaum, 2015), which means that the required tests are performed on a subset of transactions in order to generate an acceptable foundation for a conclusion about all transactions. Nevertheless, more data-driven audit approaches are gaining popularity as it provides greater coverage (since a model can check all transactions) (Earley, 2015), whereby transactions are automatically analyzed, potential errors flagged, and trends in the data can be found, which could be missed when using a sampling approach, improving the quality of audits. However, the industry continues to face challenges regarding the reliability of performance, explainability of models, and adherence to regulatory requirements.

This thesis is written in collaboration with the Dutch Central Government Audit Service (CGAS) as they actively seek solutions to overcome this emerging challenge and improve the efficiency and accuracy of financial audits. The CGAS serves as both the independent internal auditor of the Dutch government and the audit authority for the European Commission in the Netherlands (ADR, n.d.-b; Rijksoverheid, n.d.). Reporting directly to the Deputy Secretary-General and working on behalf of the ministries, the CGAS plays a vital role in examining governance, control, and accountability to ensure the quality of the government (ADR, n.d.-b, n.d.-a). Currently, the CGAS relies on a random sampling approach to select transactions for financial audits. However, there is a strong motivation to transition to a stratified sampling method facilitated by machine learning models. This transition would leverage the model's ability to check every transaction, improving the efficiency and effectiveness of audits. However, the lack of understanding regarding the performance of these models on unseen data poses a significant constraint in progressing towards this transition.
The Dutch Central Government Audit Service (CGAS) is particularly interested in obtaining an estimated confusion matrix for machine learning models applied to financial audits.

By obtaining an estimated confusion matrix, the CGAS aims to gain insights into the model's performance in terms of correctly classifying transactions as either correctly labeled or falsely labeled. By analyzing these metrics, the CGAS can identify limitations and biases, enabling them to account for the model's performance characteristics when designing a stratified sampling approach for transaction selection in audits.

This paper aims to address one of these challenges by finding a method to estimate the performance of machine learning classifier on unlabeled data. It attempts to demonstrate how reliable and accurate machine learning models for financial audits are on unseen data, making it possible to assign a statistical uncertainty to the audit statement. Therefore, the primary focus of this study revolves around the following research question:

How can the performance of machine learning models be estimated on unseen accounting data?

To conduct this research, data collected by the Dutch Central Government Audit Service (CGAS) to perform the external audit of the ministries within the Dutch central government was used. The dataset utilized in this study consists of transactions recorded during the years 2017 and 2018 in the general ledgers of two Dutch ministries. The resulting dataset comprises a total of 761,694 transactions, categorized into five distinct classes.

The remainder of this paper is structured as follows. In section 2 the literature and related work will be briefly discussed. In section 3, the dataset and its characteristics will be introduced. Section 4 describes the methodology and the approach taken. In section 5 the results of the conducted experiments are described. In section 6 the research the results will be discussed, and the research question will be answered. After this paper will be concluded by its limitations and potential areas for future research in section 7.

2. Related work

This chapter aims to provide a comprehensive understanding of the current state of research related to model performance evaluation on unseen data and the challenges and opportunities of applying machine learning models in financial audits.

2.1 Machine Learning in the Audit Domain

Machine learning (ML) has received increased attention in the audit area in recent years due to its potential to improve audit quality and efficiency (Rabade, 2022). Previous research in this area is centered primarily around two themes: detecting fraud at the financial statement level and identifying anomalies at the transaction level. Fraud detection on financial statements involves analyzing the overall financial statements to identify fraudulent activities, often based on financial ratios (Ata & Seyrek, 2009; Lin et al., 2015; Xiuguo & Shengyong, 2022). On the other hand, anomaly detection on the transaction level involves analyzing individual transactions to identify any anomalies or outliers that may indicate fraudulent activities (Vanini et al., 2023). Several studies have focused on fraud detection in financial statements by utilizing machine learning techniques that analyze text in the Management Discussion & Analysis (MD&A) sections, which was found to be a strong predictor of fraud (Purda & Skillicorn, 2015).

However, in practice, utilizing these methods is often impractical because many models have a limitation that the effectiveness must be confirmed by testing against random samples (Uijen, 2019). While this validation process is essential for ensuring the reliability and accuracy of machine learning models, obtaining labeled data in the audit environment can be particularly challenging, as financial data is often confidential and sensitive companies are cautious about using this information in production (De Roux et al., 2018). Moreover, even if these prerequisites for using sensitive data in production are met, the process of manually labeling large volumes of data can be time-consuming and resource intensive. Auditors would need to allocate significant resources to carefully annotate the data, ensuring that each entry is correctly and consistently classified (De Roux et al., 2018). To assure the quality and reliability of the labeled data, annotators must have a thorough understanding of auditing standards as well as domain knowledge.

Changes in the distribution of the data caused by a shift in expenses can cause uncertainty in the performance of models, possibly causing them to miss financial misstatements. This highlights the importance of the challenge of out-of-sample performance in machine learning for auditing. Out-of-sample performance refers to the ability of a model to accurately predict and generalize to new, unseen data that differs from the training set (Meyer-Bullerdiek, 2021). This is essential because auditors need to models to work effectively year over year on transactions and distributions, which may differ from the training data (De Roux et al., 2018). To address this problem Bao et al., (2019) attempted to create a new fraud prediction model, which aimed to minimize out-of-sample prediction error. This was done to increase the reliability of ML algorithms in auditing. However, there was no method offered to evaluate the performance of new classifiers.

2.2 Machine learning performance evaluation on unseen data

The machine learning domain aims to develop models that can effectively generalize to unseen data. However, the performance of these models on unseen data is influenced by various factors such as model complexity, data quality, and training data quantity. The model's ability to perform well on unseen data is critical for its practical application. When evaluating the model on a new dataset, common metrics like accuracy, precision, recall, and F1 score are typically employed. However, these metrics can be misleading if the training and testing data are from different distributions, making it challenging to accurately assess the model's performance (Tiittanen et al., 2019).

One widely used approach to mitigate the impact of changes in the distribution is cross-validation (CV). CV procedures involve splitting the available data into multiple subsets, training the model on a subset, and evaluating its performance on the remaining subset. By repeating this process with different subset combinations, CV provides a more reliable estimation with lower variance of the model's performance (Avula et al., 2022). However, even with CV, the estimated performance may not accurately reflect the model's performance on unseen data if the distribution of the unseen data significantly deviates from the training data, thereby making it difficult to assess the model's performance in cases where no ground truth is available.

This is particularly important in classification methods. These algorithms are designed to categorize data into specific groups based on the data's features or characteristics. To overcome this issue, the model must be re-evaluated using real-world data, however obtaining annotated test samples can be difficult and costly (Deng & Zheng, 2021). This raises the question whether model performance can be estimated on a test set without test labels or ground-truth (Botchkarev, 2018).

Another approach, developed by Schelter et al., (2020), leverages the power of unsupervised learning to extract meaningful patterns and representations. These patterns are used to measure the similarity or dissimilarity between the source and serving data distributions, enabling an estimation of the model's performance on unseen data. This approach is unique in that it relies on domain knowledge to programmatically specify typical cases of dataset shifts and data errors. By incorporating this information, a performance predictor is trained for pretrained blackbox models, which automatically raises alarms when performance drops are detected on unseen data. However, one drawback of this method is its reliance on precise and comprehensive programmatic specifications (Schelter et al., 2020). If the specifications are incomplete or inaccurate, the performance predictor may fail to detect performance drops effectively. Furthermore, the interpretability of this approach poses a challenge, as it may be difficult to explain the underlying factors contributing to performance variations.

To deal with scenarios where no ground truth (or labels) is available, domain adaptation strategies have been created (Guillory et al., 2021). Domain adaptation is a method for transferring knowledge from a source domain with labeled data to a target domain, where only unlabeled data is available (Guillory et al., 2021). However, it has been demonstrated that adapting a model between domains is challenging. Recently, a new method called Confidence-Based Performance Estimation (CBPE) has been proposed to address this challenge. CBPE utilizes the calibrated classifier's confidence scores to estimate its performance on the test set, regardless of the distribution shift between the training and test data (NannyML, n.d.). Leveraging the fact that the expected quality of the classification is contained within the class probabilities predicted by the model (Farooq et al., 2023). To anticipate a confusion matrix based on binary class probabilities, CPBE uses calibrated probability for both classes as input. The probability threshold (t) determines the boundary between classes 0 and 1 (Humphrey et al., 2022). This method has shown promising results in various applications and demonstrates its potential to in estimating the performance of classifiers on unseen data without ground truth.

2.3 Model Calibration

Calibration can be seen as an important aspect in estimating model performance, as it helps to ensure the reliability of the model's predictions by aligning the predicted probabilities with the true probabilities of the events it is predicting (Bella et al., 2010; Davis et al., 2017). Calibration serves as a mechanism to refine the output probabilities generated by a classifier, making them more representative of the actual likelihoods associated with each class. For example, items with a class probability of 0.70, should have an 70% chance of belonging to the first class (Humphrey et al., 2022). Calibration also helps to reduce the uncertainty associated with model predictions (Pasquier & Smith, 2015). By achieving proper calibration, a model can overcome issues of overconfidence or underconfidence, where the predicted probabilities may be overly optimistic or pessimistic compared to the actual probabilities. Predicting probabilities provides flexibility in interpreting them, present predictions with uncertainty, and provide more nuanced ways to evaluate the model's performance (Ferrell, 1994). For example, one may choose to interpret high probabilities as strong indications of class membership or use them to assess the level of certainty in the prediction.

However, many classifiers provide biased and poorly calibrated class probabilities, which could lead to incorrect decisions. This happens since classifiers are usually optimized for maximizing accuracy rather than producing well-calibrated probabilities. Especially for binary classifiers, but also applicable for multiclass models, is that they generate scores that are frequently interpreted as probabilities, but these scores are not real probabilities since they fail to reflect the true probability of the event of interest, making them poorly calibrated (Kuhn & Johnson, 2013; Gokcesu & Gokcesu, 2021).

To address calibration issues, various calibration methods have been developed. Platt scaling, isotonic regression and temperature scaling are three widely used techniques that aim to refine the predicted probabilities and bring them closer to the true probabilities, enhancing the model's calibration and reliability (Zadrozny & Elkan, 2002). These methods are often visualized in a calibration curve. These curves visualize the relationship between predicted probabilities and observed frequencies of events.

Metrics such as log-loss, Brier score, and Expected Calibration Error (ECE) are widely used to assess the calibration of probabilistic models (Trottini et al., 2020). Log-loss quantifies the average negative logarithm of the predicted probabilities, penalizing models for assigning low probabilities to true events. Brier score measures the mean squared difference between predicted probabilities and true probabilities, providing a measure of overall calibration accuracy. Additionally, ECE evaluates the calibration of predicted probabilities across different probability intervals or bins, providing insights into the model's calibration across the entire probability range (Nixon et al., 2020).

2.4 Drift

Drift refers to a critical issue in machine learning where the assumption of a stable data distribution over time is violated. This assumption is often implicit in typical machine learning approaches, leading to potential problems when the data distribution changes over time, resulting in what is known as "drift." Drift can cause a phenomenon called "silent failure," where the model's predictions degrade in quality compared to its initial validation performance. There are two significant types of drift that can significantly impact machine learning models: data drift and concept drift (Bennett et al., 2022; El-Hay & Yanover, 2022).

These terms are often used interchangeably in the literature, but in this paper, a distinction will be made to highlight their differences. In this work, concept drift is defined as a shift in the relationship between the input variables and the target variable, so that a new dataset with statistically equal characteristics receives different labels (Humphrey et al., 2022). It occurs when a new dataset, despite having statistically similar characteristics to the original data, receives different labels. In other words, the underlying concept that was learned by the model changed over time. Data drift also known as covariate shift or population shift, can be defined as an occurrence in which the overall data distribution of inputs variables shifts, without changing the mapping between the features and labels. In other words, the relationship between the features and labels remains the same, but the statistical properties of the input data change (Sugiyama, 2016).

Aside from drift, another challenge is out-of-distribution (OOD) data. OOD data refers to samples that differ significantly from the training data and are not adequately represented by the model's learned distribution. When faced with OOD data, machine learning models tend to provide unreliable predictions, as they have not been trained on such samples during the learning process. OOD data can arise in various scenarios, such as encountering novel or unexpected examples, data collected from different sources, or samples from previously unobserved contexts (Botchkarev, 2018). The lack of exposure to OOD data during training makes it difficult for models to generalize effectively and accurately predict the outcomes for these unfamiliar samples. This is one of the reasons it can be difficult to apply transfer learning as a model may not generalize well to the new task, leading to reduced accuracy (Hsu et al., 2020).

3. Dataset Description

The dataset used in this study consists of transactions recorded in the general ledgers of two Dutch ministries during 2017 and 2018. Uijen (2019) previously conducted a separate research study in which data from two Dutch ministries were merged. The complete dataset was split in three subsets for this analysis. Table 1 provides an overview of the dataset, which includes the complete dataset (761,694 transactions in total) as well as its subsets: the training set, calibration set, and test set. The data is randomly split into a 70:15:15 ratio. Notably, there was no missing data in the dataset. It shows the frequency and percentage distribution of the various class labels within each subset. The class labels represent five distinct categories of expenses, such as Staff Costs, Purchase of Goods and Services, Program Expenses, Depreciation & Impairment, and Interest Costs. Importantly, the class label distributions in each subset, including the training, calibration, and test sets, are highly comparable. This similarity ensures that the subsets are consistent and comparable, allowing for reliable analysis and evaluation.

Table 1: Distribution of transactions across classes and subsets

	Entire Dataset		Training Set		Calibration Set		Test Set	
Class Label	Frequency	%	Frequency	%	Frequency	%	Frequency	%
Staff Costs	331,822	43.6	232,435	43.6	49792	43.6	49595	43.4
Purchase of Goods and Services	284,275	37.3	199,222	37.4	42488	37.2	42565	37.3
Program Expenses	139,045	18.3	96929	18.2	20991	18.4	21125	18.5
Depreciation and Impairment	4195	0.6	2918	0.5	647	0.6	630	0.6
Interest Costs	2357	0.3	1681	0.3	337	0.3	339	0.3
Total	761,694	100	533,185	100	114,255	100	114,254	100

Note: the percentages were rounded to the first decimal

In the context of ethical and legal considerations, it is important to mention that the data was gathered by the CGAS. The CGAS operates in compliance with legal and regulatory frameworks governing data privacy and confidentiality. It is crucial to emphasize the ethical and legal considerations associated with the utilization of this dataset, which contains transactions from the general ledgers of two Dutch ministries. Due to the sensitive nature of the data, strict confidentiality measures are in place, including the presence of a non-disclosure agreement (NDA). Access to the data is limited and can only be granted through a ministry-provided virtual machine or a government computer, ensuring the integrity and confidentiality of the information.

An ethical implication of using general ledgers for research purposes is the potential presence of personal information within the dataset. General ledgers often contain financial transactions that may include personal identifiers, such as names, addresses, or other sensitive details. In addition, the presence of a non-disclosure agreement (NDA) poses an ethical limitation in terms of data transparency and openness. The NDA restricts the researcher from freely sharing or making the data open source, potentially limiting the ability of the wider research community to access and validate the findings[1].

[1] The dataset is maintained by the CGAS. The code is accessible upon request and will also be made publicly available on GitHub in the future. For inquiries and access requests, please contact me at alexessayan+thesis@gmail.com.

4. Methodology

This chapter provides a comprehensive description of the methodology employed in this paper. The methodology serves as the foundation for conducting the research, guiding the processes of data collection, analysis, and interpretation. By outlining the approach taken, this chapter aims to offer transparency and clarity regarding the methods used to address the research objective.

4.1 Model Description

This research builds upon models developed by Uijen (2019), specifically utilizing the CNN architecture due to its higher performance as a probabilistic multiclass classifier. While the used approach can be applied to various classifiers, the emphasis is on error detection based on text inputs, particularly transaction descriptions. To represent the text descriptions, the model uses concatenated word vectors (a numerical representation of words) as input, with each vector corresponding to a specific word. Leveraging the CNN architecture, originally designed for computer vision but effective in NLP (Yu et al., 2021; Li et al., 2022), the model extracts position-invariant and local features from the text.

The CNN-based text representation involves convolution and pooling operations, playing vital roles in feature extraction. Convolution applies filters to capture N-gram features at various points in sentences, resulting in feature maps. Pooling operations capture short and long-term relationships within the sentence. The model handles numerous N-grams at the same time through using filters of different lengths, improving text representation and classification accuracy.

The model architecture includes convolutional, dropout, and fully connected layers. The input consists of three transaction descriptions, transformed into continuous vectors through embedding layers. Convolutional layers apply filters to capture N-gram features, followed by dropout layers to prevent overfitting. Flattening and concatenation layers combine the extracted features, and fully connected layers capture higher-level relationships. The dense output layer with softmax activation generates the probability distribution over the five class labels (as shown in Table 1), indicating the probabilities that a given transaction belongs to each class. As a result, for each transaction, the model generates five probabilities denoting the possibility of it being classified into each class.

4.2 Probability Calibration

First, the model is used to make predictions on the different subsets of the data. The output of this model (the array of probabilities) forms the uncalibrated probabilities. Different kind of probability calibration methods have been applied to the uncalibrated probabilities. After the application of the different calibration methods, the calibrated probabilities will be compared with the uncalibrated probabilities by the scoring rules described later in this chapter. The methods applied in this study are Platt Scaling, Isotonic Scaling, and Temperature Scaling since they have been extensively studied and are commonly used (Kumar et al., 2020; Tabacof & Costabello, 2020).

Platt Scaling was the first method that was utilized to attempt to improve the calibration of the model. This method was chosen for its simplicity and performance, making it easy to interpret while effectively calibrating the probabilities (Böken, 2021). Platt Scaling was applied using the CalibratedClassifierCV function of Scikit-Learn. Logistic regression with the lbfgs solver is used to fit a calibration classifier, with a maximum of 1000 iterations. The calibration process includes ten rounds of cross-validation, ensuring robustness. The sigmoid method is employed for calibration, attempting to align the predicted probabilities with the true probabilities. The calibrated_clf object represents the calibrator, fitted on the predicted probabilities and truth labels. Since this method was applied in

a multiclass scenario, Platt scaling was extended using a one-versus-all strategy (CalibratedClassifierCV does this by default), applying calibration individually to each binary classifier (Johansson et al., 2021). By adjusting the predicted probabilities for each class, the resulting calibrated multiclass probabilities provide more accurate estimates and enhance the calibration of the multiclass classification model.

The second calibration method utilized was Isotonic Scaling. This strength of this method lies in its ability to effectively fit non-parametric data. However, this can become a weakness when data is limited, as it has been shown that it is more prone to overfitting (Niculescu-Mizil & Caruana, 2005). Scikit-learn's IsotonicRegression was used for this purpose. The textual labels were translated into numeric values before fitting the isotonic regressor to the calibration set using label encoding. For this conversion, the scikit-learn LabelEncoder class was used. Using a loop across each class, the isotonic calibration was then applied to the calibration set. It is important to note that isotonic Scaling, as applied in this manner, follows a one-versus-all strategy in the context of this multiclass calibration, similar to how Platt Scaling was applied (Song et al., 2018). By treating each class as a binary classification problem against the rest of the classes, the Isotonic Scaling method adjusts the predicted probabilities to improve calibration and align them with the true probabilities for each class. To handle any out-of-bounds issues, an IsotonicRegression object was formed within the loop with the 'out_of_bounds' option set to 'clip'. Out-of-bounds issues occur when predicted probabilities or values fall outside the expected or valid range. The calibration probabilities for the specified class, as well as the matching encoded labels obtained from label encoding, were used to fit the isotonic regressor. This procedure generated calibrated probabilities for the calibration set.

The third calibration method applied in this context is Temperature Scaling. This method adjusts the temperature parameter to align the predicted probabilities with the true probabilities, thus attempting to improve the calibration of the predicted probabilities (Kull, Perello-Nieto, et al., 2019). The primary notion behind temperature scaling is that greater temperatures increase the entropy of predicted probability, spreading them out and making them less confident. Leading to a more uniform distribution of probabilities. Lower temperatures, on the other hand, sharpen the probability distribution, allowing the classifier to make more accurate predictions (Guo et al., 2017).

The temperature_scale function takes as input the predicted probabilities and a temperature value. It scales the probabilities by exponentiating them with the temperature and normalizing them to obtain a calibrated probability distribution. This ensures that the probabilities sum up to 1 for each instance. To determine the best temperature value for calibration, a range of temperature values is defined using np.linspace, spanning from 0.1 to 10.0 with 50 equally spaced values (Joy et al., 2022). The goal is to find the temperature that yields the lowest log loss when comparing the calibrated probabilities to the calibration labels. The calibration set's best temperature is afterwards used on the test set to adjust the sharpness of the softmax function (Kull, Perelló Nieto, et al., 2019).

4.3 Scoring rules

The calibration will be scored at three metrics the Expected Calibration Error (ECE), Log Loss and Brier score. Where the model scoring best at least two of the three measures will be picked for further analysis, considering there won't be a significant drop in accuracy. The three measures are used simultaneously to obtain a more comprehensive evaluation of the calibration. But as calibration could have a costly impact on accuracy due to a variety of reasons, including overconfidence, underfitting, and data imbalance accuracy will also be assessed (Tabacof & Costabello, 2020).

4.3.1 Expected Calibration Error

Expected Calibration Error (ECE) is a frequently used scoring rule to evaluate the calibration of predicted probabilities in classification models. It quantifies the difference between the expected probability and their empirical probabilities. This is accomplished by dividing the predicted probability into evenly spaced bins and computing the average absolute difference between the

predicted and empirical probabilities inside each bin. In this case there is chosen for a ECE with 10 bins. A lower ECE number indicates greater calibration, which makes it intuitive to interpret (Johansson et al., 2021). ECE for multiclass calibration can be calculated as follows:

$$ECE = \frac{1}{K}\sum_{k=1}^{K}\sum_{b=1}^{B}\frac{n_{bk}}{N}|acc(b,k) - conf(b,k)|$$

The ECE is first calculated for each class (k) accuracy and confidence of bin b for class label k are represented by $acc(b,k)$ and $conf(b,k)$ respectively; n_{bk} is the number of predictions in bin b for class label k; and N is the total number of data points. The absolute difference between $acc(b,k)$ and $conf(b,k)$ is computed for each bin and class to determine the ECE (Nixon et al., 2020). This difference is then multiplied by the fraction (n_{bk}/N), which accounts for each bin's contribution to the overall ECE score. The values obtained are summed across all bins and classes. Finally, the summed values are averaged across the K classes to provide an overall assessment of the classification model's calibration performance.

One of the limitations of ECE is that it focuses on the calibration of predicted probabilities within discrete bins. This approach might not capture the calibration performance accurately in cases where the predicted probabilities exhibit complex or nonlinear relationships across the probability range. The ECE was calculated for each class within each method and evaluated the calibration of predicted probabilities, returning the average ECE per method for comparison. Predicting the overall class distribution regardless of the given case, is a simple way of obtaining precisely calibrated probabilities (Kull, Perelló Nieto, et al., 2019). As a result, there is chosen to incorporate other measures, namely Log Loss and the Brier score.

4.3.2 Log Loss

Log loss is another commonly used measure to check calibration as it's a proper scoring rule that incentives the model to output well calibrated probabilities. The log loss penalizes the model for being both overconfident and underconfident, which is important in classification problems where the predicted probabilities need to match the true probabilities of the events being classified (Lucena, 2018). The downside of using log loss is that it can be difficult to optimize directly, as it is not a convex function. It is also sensitive to class imbalance, meaning that it can be biased towards the majority class in imbalanced datasets (Ben-Yishai & Ordentlich, 2021). The Log Loss can be computed as follows: $L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(P_{ij})$

where N is the number of events being classified, *M* is the number of classes, y_{ij} is an indicator variable that is 1 if event i belongs to class j and 0 otherwise, and p_{ij} is the predicted probability of event i belonging to class j.

4.3.3 Brier score

The last measure that is used to compare the different calibration method is the Brier score. The main advantage of this method is that it is less sensitive to class imbalance than log loss, meaning it is more appropriate for imbalanced data sets. The limitation is that the Brier score can be less sensitive to small changes in the predicted probability (Mosquera et al., 2022). The Brier score is defined as:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(P_i - o_i)^2$$

where N is again the number of events being classified, p_i is the predicted probability of the event being in the positive class (i.e., having label 1), and o_i is the true label of the i_{th} event (either 0 or 1). The Brier score measures the mean squared difference between the predicted probabilities and the true labels, and ranges from 0 (perfect calibration) to 1 (worst calibration) (Bellinger et al., 2020). For

each method the Brier score was calculated per class using the predicted probabilities and the actual category labels. Then the average brier score (over all classes) for per each method was assessed.

4.4 Performance Estimation

After the best calibrator has been chosen, this will be applied to the test set and then we will look at how the performance can be estimated without truth labels. These estimation methods will be compared to the same test set with truth labels. To estimate the performance of the chosen model, the open-source Python module NannyML will be used to perform a Confidence-Based Performance Estimation (CBPE). This approach may offer predicted performance of a classifier in the lack of ground truth, even in the midst of data drift (Farooq et al., 2023). In essence, CBPE utilizes the anticipated performance of the classifier by inputting its calibrated scores (Humphrey et al., 2022). Therefore, the probabilities calibrated by the best calibrator will be utilized. In addition to implementing the standard NannyML algorithm, independent code was developed to calculate specific performance metrics, using the same method but without relying on this package to ensure transparency. To complete this research within set time, the assumption is made that there is no concept drift.

In multiclass models, both prediction labels and probability estimates are provided for each class. This means that for a model with 5 classes, the output consists of the predicted class along with the corresponding probabilities for all 5 classes. The class with the highest probability will be identified as the predicted class using the argmax function. One commonly used metric in multiclass classification is macro-averaged precision, which can be estimated using the CBPE method as follows:

Firstly, precision is estimated for each class separately, treating it as a binary classification problem. The multiclass prediction vector $\hat{y}$ is transformed into a binary vector $\hat{y}_k$ relevant to class k, accompanied by the corresponding predicted probabilities $\hat{p}_k$. The precision$_k$ for class k is calculated using the precision function, considering $\hat{y}_k$ as the predicted labels and $\hat{p}_k$ as the corresponding probabilities. The binary vector $\hat{y}_k$ is constructed by setting $\hat{y}_k, j$ as 1 if the predicted class $\hat{y}$ for observation j is equal to k, and as 0 otherwise. After estimating the precision$_k$ for each class, the macro-averaged precision is obtained by averaging across all classes, yielding the overall precision for the multiclass classification model.

The estimation of recall, F1 score, specificity, and one-versus-rest ROC AUC follows a similar approach using the predicted labels and probabilities (Guillory et al., 2021). Average Confidence (AC) or multiclass accuracy could be estimated as the mean of predicted probabilities corresponding to the predicted classes. As the classifier is argmax calibrated on the target, then the average confidence accuracy is expected to be equal to accuracy of the classifier (Garg et al., 2022) .

Since the auditors are primarily interested in minimizing the number of false positives, meaning a transaction is classified as correct, but is incorrectly categorized. For this the measure false positive rate can be used best. To calculate the False Positive Rate (FPR) for each class, the algorithm follows these steps: First, it obtains the probability estimates $\hat{p}$ for each class from the multiclass classification model. Then, it determines the predicted label $\hat{y}$ for each instance by selecting the class with the highest probability estimate using the argmax operation. A threshold value is set to determine whether an instance is classified as positive or negative for a specific class. Next, the algorithm calculates the False Positives (FP) for each class by considering instances that are predicted as positive for that class but do not exceed the threshold (Garg et al., 2022). For class k, the False Positives (FP$_k$) are calculated as the sum of 1 if ($\hat{y}_k$ = 1 and $\hat{p}_k$ < threshold) for all instances and 0 otherwise. Furthermore, the algorithm estimates the True Negatives (TN) for a specific class k by counting the instances that are classified as negative for that class, i.e., instances where the predicted label is not k. For class k, the True Negatives (TN$_k$) are calculated as the sum of 1 if ($\hat{y}_k \neq k$) for all instances and 0 otherwise. Finally, the False Positive Rate (FPR) for each class k is calculated by

dividing the False Positives (FP_k) by the sum of False Positives (FP_k) and estimated True Negatives (TN_k). Thus, for class k, the False Positive Rate (FPR_k) is calculated as $FPR_k = \frac{FP_k}{(FP_k+TN_k)}$.

5. Results

The original research objective aimed to investigate the estimation of a confusion matrix in the context of multiclass classification. However, this pursuit has proven to be exceptionally complex. The complexity lies in accurately determining the components of the confusion matrix without access to the ground truth labels.

In binary classification, adjusting the classification threshold is an useful technique to manage the balance between false positives and true negatives (Miller et al., 2018). By manipulating the threshold, instances can be categorized as positive or negative based on the classifier's output probability or score, leading to a trade-off between precision and recall and influencing the resulting confusion matrix.

However, in a multiclass scenario, the interpretation of the confusion matrix heavily depends on the context of each individual class being considered as the positive class. As each class is treated as a distinct positive class against the remaining classes considered as the negative class in a one-versus-all approach. For example, determining whether a prediction is a false positive or a true negative depends on the specific positive class being considered. This becomes particularly challenging as instances are misclassified as positive for one class while belonging to another positive class, leading to ambiguity between false positives and true negatives. Consequently, adjusting a threshold alone would not be sufficient to address the issue of appropriately categorizing instances in a multiclass context. Although the objective of estimating a reliable confusion matrix becomes increasingly elusive, alternative methods, rather than relying on the multiclass confusion matrix, have been explored to estimate the performance of the classifier.

One of these methods to estimate the performance of a classifier without truth labeled is the Confidence Based Performance Estimation as described in the methodology section. One of its assumptions is that the probabilities used as input are well calibrated. For this reason, probability calibration was assessed using different methods. The methods evaluated include Uncalibrated, Platt Scaling, Temperature Scaling, and Isotonic Scaling. The performance metrics used in this analysis include Log Loss, Expected Calibration Error (ECE), Mean Brier Score, and Accuracy. The performance of the different calibration methods is presented in Table 2.

Table 2: The performance on the calibraton metrics per calibration method

Method	Log Loss	ECE	Mean Brier Score	Accuracy
Uncalibrated	0.486831	0.092915	0.155729	0.810046
Platt Scaling	0.535776	0.124329	0.141356	0.808401
Temperature Scaling	0.598169	0.121455	0.122080	0.810046
Isotonic Scaling	0.544036	0.103306	0.150930	0.810046

The uncalibrated method represents the baseline performance, where the predicted probabilities are not adjusted. It achieved a Log Loss of 0.486831, which measures the accuracy of probability predictions. The Expected Calibration Error (ECE) of 0.092915 indicates the average difference between the predicted probabilities and their true values. The Mean Brier Score of 0.155729 represents the overall accuracy and reliability of the predicted probabilities, while the Accuracy of 0.810046 measures the overall correctness of the predictions.

Platt Scaling yielded a slightly higher Log Loss (0.535776) and ECE (0.124329) compared to the Uncalibrated probabilities. However, it achieved a slightly lower Mean Brier Score (0.141356),

indicating a slight improvement in overall reliability. The Accuracy was also slightly influenced, resulting in a score of 0.808401. Overall, Platt Scaling exhibited a slightly worse performance compared to the Uncalibrated probabilities, but the difference was still within a comparable range.

Temperature Scaling adjusts the temperature parameter. The temperature was chosen by finding the best temperature on the calibration set. The best temperature was rounded to three decimals 0.302. It resulted in a higher Log Loss of 0.598169 and ECE of 0.121455 compared to the uncalibrated probabilities. However, it achieved the lowest Mean Brier Score of 0.122080 among all the methods, indicating slightly improved accuracy and reliability. The Accuracy of 0.810046 suggests that it performs similarly to Uncalibrated in terms of overall correctness.

Isotonic Scaling, the only non-parametric calibration method used to calibrate the predicted probabilities. It achieved similar scores to the other methods, with a Log Loss of 0.544036, an ECE of 0.103306, and a Mean Brier Score of 0.150930. The Accuracy was the same with the other methods, except for Platt Scaling.

As the uncalibrated probabilities demonstrated the best calibration performance, as reflected in the Log Loss and ECE metrics. They were selected to be used throughout the remainder of this paper. It is worth noting that the presented scores align with the findings from the reliability graph (Figure 1).

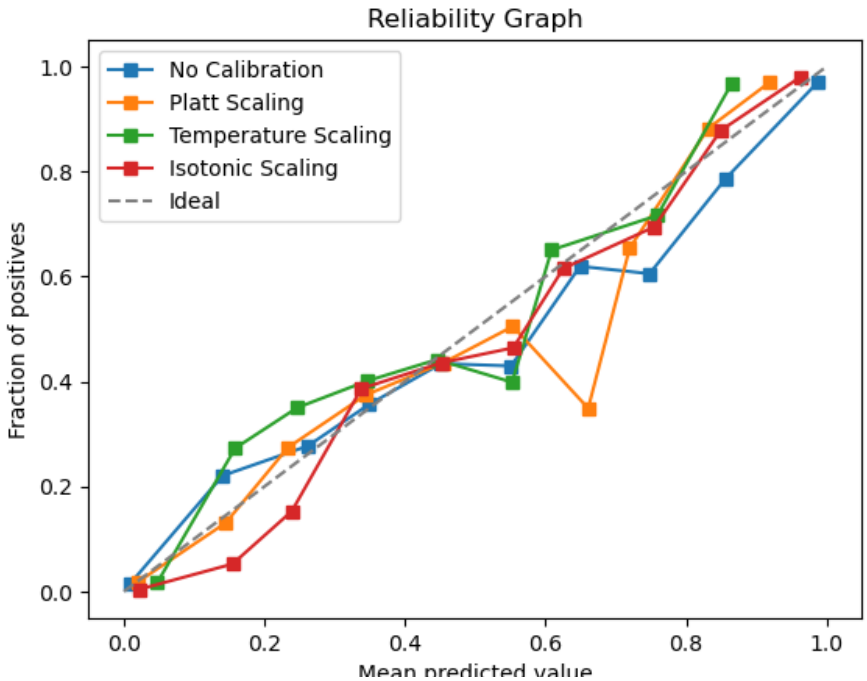

Figure 1: Reliability Graph of the calibration methods

After the calibration the performance of the classification model was evaluated using both ground truth values and estimated values. Table 3 presents the results based on the ground truth values, while Table 4 displays the estimated values. In terms of precision, the estimated values in Table 4 generally align well with the ground truth values from Table 3, even though they were slightly more optimistic. For example, in the class Staff Costs the estimated precision was slightly higher (0.7853) compared to the ground truth precision of 0.77. The most notable difference was for the class Purchase of Goods and Services where the truth value was 0.85 and the estimated was 0.9158. But this could still be considered as within comparable range. The estimations for precision where within 6% of the actual values.

Table 3: Classification report with ground truth

Classification Report true values				
Class label	Precision	Recall	F1-Score	Support
Staff Costs	0.77	0.89	0.83	49595
Purchase of Goods and Services	0.85	0.77	0.81	42565
Program Expenses	0.83	0.74	0.79	21125
Depreciation and Impairment	0.98	0.13	0.23	630
Interest Costs	0.00	0.00	0.00	339
Accuracy			0.81	114254
Macro average	0.69	0.51	0.53	114254
Weighted average	0.81	0.81	0.81	114254

When examining the recall values, the estimated scores seem to be higher than the true values across classes. This aligns with the observations made for precision, indicating an overly optimistic estimation of the model performance in terms of class identification. However, the estimations despite being higher than the true values in Table 3, still demonstrate a reasonable level of similarity. As all the classes, except Program Expenses, fall within an approximate range of 3% compared to the truth values. The estimation for Program Expenses stands out with a significantly larger margin of error, approximately 10%.

Table 4: Estimated classification report

Estimated Classification Report				
Class label	Precision	Recall	F1-Score	Support
Staff Costs	0.7853	0.9269	0.8502	56887
Purchase of Goods and Services	0.9158	0.7719	0.8377	38470
Program Expenses	0.8707	0.8441	0.8572	18814
Depreciation and Impairment	0.9631	0.1346	0.2362	83
Interest Costs	0.0000	0.0000	0.0000	0
Average Confidence Accuracy			0.8434	114254
Macro average	0.7070	0.5355	0.5563	114254
Weighted average	0.8434	0.8605	0.8467	114254

The estimated F1-scores tend to be higher across classes compared to the true F1-scores. Most classes have a relatively small margin of error, falling within an approximate range of 3% deviation from the true values. This suggests that the model's performance in terms of class identification is reasonably consistent between the estimated and true values. As with precision and recall, the estimations where more optimistic about the classifier's performance. The program expenses, like recall, had the highest margin of error of approximately 6%.

The average confidence accuracy, which serves as an estimated accuracy of the classifier, was calculated to be 0.8434. A comparison with the real accuracy of 0.81 reveals a slight positive margin of error, approximately 3%. This finding suggests that comparable to other confidence-based measures, the average confidence accuracy tends to overestimate the true accuracy of the classifier.

The macro average calculates the average performance across classes with equal weight, while the weighted average considers the average performance with class-specific weights based on their support or sample size. For the macro average, Table 3 reports a precision, recall, and F1-score of 0.69, 0.51, and 0.53, respectively. In Table 4, the estimated values for the macro average are 0.7070, 0.5355, and 0.5563, respectively. Showing that the estimations are yet again more optimistic, but all are within the 3% margin of error, in line with the individual class scores. The weighted average

followed a similar pattern, but with a somewhat wider error margin. This difference can be attributed to the higher weight assigned to the class Program Expenses within this metric in comparison to the macro average, as this class had a relatively high variance.

Because this paper focuses on the audit domain, where false positives are of particular concern, the false positive rate (FPR) was estimated based. This estimation was based on the probabilities and a threshold. This involved identifying instances predicted as positive for each class and those instances where the predicted probability fell below a predetermined threshold as described in the methodology.

To address the challenges posed by class imbalance and the potential costs associated with false positives, a threshold higher than the conventional value of 0.5 was selected in this study (Esposito et al., 2021). This higher threshold ensured a more conservative approach in classifying instances as positive, thereby reducing the likelihood of false positives. In this study, it was decided to use a single threshold for all classes, assuming a roughly equal importance in correctly classifying each class (Balayla, 2021). Specifically, a threshold of 0.62 was chosen to minimize the occurrence of false positives, considering the unique requirements and considerations of the audit domain. The threshold value of 0.62 was chosen considering the desire to minimize false positives, and it appeared to be a suitable value after examining possible threshold values on the calibration set.

Table 5 presents the comparison of the false positive rates (FPR) with the ground truth values and the estimated FPR values calculated using the method described above. For the Staff Costs class, the ground truth FPR is 7.92%, while the estimated FPR is 6.41%. In the Purchase of Goods and Services class, the ground truth FPR is 3.36%, and the estimated FPR is 3.18%. The Program Expenses class exhibits a ground truth FPR of 19.95%, whereas the estimated FPR is significantly lower at 3.66%. For the Depreciation and Impairment class, due to a limited number of samples (only 83), the ground truth FPR could not be accurately measured. Similarly, for Interest Costs, there were no samples identified as positive, resulting in zero FPR for both the ground truth and the estimated values (approximately). This table shows that the estimated values come in a similar margin of error to the other performance metrics in the classification report. However, there is a serious mismatch for the Program Expenses class. This disparity could be due to the model's inability to reliably predict Program Expenses. When a model has limited discriminatory power to capture the distinctive characteristics of a specific class, it tends to generate more false positive predictions, resulting in a higher FPR.

Table 5: The False Positive rate and the Estimate False Positive rate per class

False Positive Rate and Estimated False Positive rate per class

Class Label	False Positive Rate	Estimated False Positive Rate
Staff Costs	0.0792	0.0641
Purchase of Goods and Services	0.0336	0.0318
Program Expenses	0.1995	0.0366
Depreciation and Impairment	0.0000	0.0000
Interest Costs	0.0000	0.0001

Furthermore, the Nannyml library has been tested, which is designed to estimate performance and detect data drift without relying on ground truth or true values. However, during testing, it was discovered that the library did not adequately reflect the classifier's true performance.
This attempt yielded no meaningful results. It is important to note that the primary goal of this paper did not involve detecting data drift. Nonetheless, it was discovered that the Nannyml library struggled to establish appropriate drift detection thresholds and failed to generate reliable confidence bands around the classifier's accuracy or F1 score. Additionally, even when the

imbalanced classes were removed, the library still performed poorly. Due to time constraints and a lack of significant discoveries, it was decided not to incorporate the results from this library in this study.

6. Discussion and Conclusion

In this section, the research question will be recapped, followed by an overview of the research methodology. Furthermore, the key findings of the study will be summarized.

The primary research question of this study was: *"How can the performance of machine learning models be estimated on unseen accounting data?"* Despite the limitations in calculating a multiclass confusion matrix without ground truth, this paper focused on estimating performance metrics that are relevant to auditors, particularly false positives, to assess the feasibility of an informed sample. For the estimation of performance metrics and to address the research question, the Confidence Based Performance Estimation (CBPE) methodology was used. First the calibration of the model was assessed, and attempts were made to improve it. Subsequently, the CBPE method was used to estimate different performance metrics, including, F1-score, and Average Confidence Accuracy, among others. Additionally, the False Positive Rate for each class was also estimated. By determining the predicted labels and setting appropriate threshold values, we were able to classify instances as positive or negative.

The findings of the calibration methods revealed that the uncalibrated model exhibited the best performance across different calibration metrics: Log Loss, ECE, Mean Brier Score, and Accuracy. While the uncalibrated model achieved better results in terms of these metrics, it should be noted that the calibration methods were primarily focused on improving the calibration of the probability estimates rather than optimizing for performance. The Log Loss score of 0.486831 indicates good calibration, with low negative logarithm values for predicted probabilities. The ECE score of 0.092915 suggests that the predicted probabilities are relatively close to the actual frequencies of events, while the Mean Brier Score of 0.155729 indicates small, squared differences between predicted and observed probabilities. This indicates that the predicted probabilities are well-aligned with the true probabilities (Kumar et al., 2020). Surprisingly, the probabilities from the model outperformed the calibration methods, which further indicates that the original probability estimates were already well-calibrated.

The CBPE method consistently provided close estimations of the real performance across various metrics, with an error range of approximately 3%, consistently on the optimistic side. This pattern could be attributed to the presence of a calibration error, as overconfidence often arises from such discrepancies (Guillory et al., 2021). However, it is worth noting that the recall for the class Program Expenses deviated significantly from the other estimations, which were closer to 10% deviation. Similarly, when estimating the False Positive Rate (FPR), the estimates were generally close, except for the class Program Expenses, where there was a notable discrepancy. This discrepancy can be attributed to the model itself, which exhibited a high number of false positives for this class, not accurately reflected in the estimations. As a result, the observed gap in the Program Expenses raises concerns about the model's tendency of producing false positives, requiring further examination of the model. It is important to acknowledge that the reliability of the estimations may not accurately reflect the underlying FPR if the model's performance for a specific class is suboptimal.

The standard graph implementation in NannyML for performance and drift detection without ground truth did not yield satisfactory results, possibly due to random data splitting instead of a chronological order or the class imbalance. Additionally, achieving an accuracy of at least 90% may be necessary to improve performance in future attempts (NannyML, n.d.).

The study's findings have practical implications beyond auditors, offering an approach to estimate machine learning model performance in the absence of ground truth. Specifically, for auditors, this method provides insights on the feasibility and usability of employing machine learning models in their audit procedures. By evaluating the performance estimations, auditors can measure the potential benefits and determine the extent to which they can integrate machine learning models. This enables auditors to move towards a more informed sampling approach, leveraging data-driven insights rather than relying solely on random sampling methods.

In conclusion, this study addressed the primary research question: "*How can the performance of machine learning models be estimated on unseen accounting data?"* By utilizing the Confidence Based Performance Estimation (CBPE) methodology, we focused on estimating relevant performance metrics for the auditors, particularly false positives. The findings highlighted the effectiveness of the CBPE method in providing close estimations of performance metrics, while also emphasizing the importance of acknowledging discrepancies in specific classes. Furthermore, this study offers practical implications for auditors in evaluating the integration of machine learning models in audit procedures and promoting data-driven insights for a more informed sampling approach.

7. Limitation & future research

When observing the progression of this study, constraints related to time, scope, bias are relevant to consider. The reliability of the probability estimates used in the study is determined by the performance of the underlying classifier. Furthermore, the study is dependent on data checked by auditors, which may be susceptible to human error, bringing the chance of inaccuracies being reflected in the model (Dechow et al., 2011).

This study did not specifically focus on calibration; however, it utilized binary calibration methods extended in a one-versus-all manner. This presents an opportunity for future research to explore the possibilities of multiclass calibration. The method's usefulness and adaptability can be improved by including multiclass calibration. For instance, in the case of temperature scaling, a potential extension could involve setting separate temperature values for each class, allowing for more fine-grained calibration specific to individual class characteristics.

The method CBPE method has limitations in terms of its binary transformation approach for multiclass classification, potentially oversimplifying the complexities and interdependencies between classes. Additionally, in this paper, class imbalance is not addressed. Future study might investigate methodologies that go beyond binary transformation for multiclass classification, as well as dealing with class imbalance in the context of these estimations.

Furthermore, the model produced many false positives, particularly for Program Expenses, influencing the estimations. Investigating the causes of false positives and determining why the model performed poorly in detecting Program Expenses, as well as investigating possible improvements could be an area for future research. Another factor to consider is the specified threshold, which was set at 0.62 for all classes. To improve estimations, future research could investigate alternative methods for threshold selection, such as data-driven approaches or class-specific thresholds. This could help balance the trade-off between false positives and false negatives based on the specific characteristics for each class, to achieve better estimations.

Additionally, the NannyML library used in this study did not provide any insights. To enhance the analysis, future research could consider incorporating a time element by segmenting the data into specific periods instead of relying on random splits. This temporal approach has the potential to improve the performance of the library, particularly in situations where the drift in data distribution over time is a critical factor. Additionally, investigating methods for real-time monitoring of model performance would be valuable, as the current approach primarily focused on static datasets.

Lastly, future research could address the impact of data and concept drift on the performance of machine learning models and this estimation methods put forward in this research, particularly in domains such as auditing where errors can have significant consequences. Understanding and mitigating the challenges associated with drift will improve the model's robustness and applicability in production.

8. Bibliography

Appelbaum, D. (2015). *SECURING BIG DATA PROVENANCE FOR AUDITORS: THE BIG DATA PROVENANCE BLACK BOX*. *13*. https://doi.org/10.5748/9788599693117-12CONTECSI/PS-2933

Arbet, J., Brokamp, C., Meinzen-Derr, J., Trinkley, K. E., & Spratt, H. M. (2020). Lessons and tips for designing a machine learning study using EHR data. *Journal of Clinical and Translational Science*, *5*(1), e21. https://doi.org/10.1017/cts.2020.513

Avula, N. V. S., Veesam, S. K., Behera, S., & Balasubramanian, S. (2022). *Building Robust Machine Learning Models for Small Chemical Science Data: The Case of Shear Viscosity* (arXiv:2208.10784). arXiv. https://doi.org/10.48550/arXiv.2208.10784

Balayla, J. (2021). *Prevalence Threshold and bounds in the Accuracy of Binary Classification Systems* (arXiv:2112.13289). arXiv. http://arxiv.org/abs/2112.13289

Bao, Y., Ke, B., Li, B., Yu, Y., & Zhang, J. (2019). Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach. *Journal of Accounting Research*, *58*. https://doi.org/10.1111/1475-679X.12292

Bella, A., Ferri, C., Hernández-Orallo, J., & Ramírez-Quintana, M. J. (2010). Calibration of Machine Learning Models. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* (pp. 128–146). IGI Global. https://doi.org/10.4018/978-1-60566-766-9.ch006

Bellinger, C., Corizzo, R., & Japkowicz, N. (2020). *ReMix: Calibrated Resampling for Class Imbalance in Deep learning* (arXiv:2012.02312). arXiv. https://doi.org/10.48550/arXiv.2012.02312

Bennett, M., Balusu, J., Hayes, K., & Kleczyk, E. J. (2022). *The Silent Problem—Machine Learning Model Failure—How to Diagnose and Fix Ailing Machine Learning Models* (arXiv:2204.10227). arXiv. https://doi.org/10.48550/arXiv.2204.10227

Ben-Yishai, A., & Ordentlich, O. (2021, February 16). *Constructing Multiclass Classifiers using Binary Classifiers Under Log-Loss*. ArXiv.Org. https://arxiv.org/abs/2102.08184v2

Böken, B. (2021). On the appropriateness of Platt scaling in classifier calibration. *Information Systems*, *95*, 101641. https://doi.org/10.1016/j.is.2020.101641

Botchkarev, A. (2018). *Evaluating Performance of Regression Machine Learning Models Using Multiple Error Metrics in Azure Machine Learning Studio* (SSRN Scholarly Paper No. 3177507). https://doi.org/10.2139/ssrn.3177507

Cinà, A. E., Grosse, K., Demontis, A., Vascon, S., Zellinger, W., Moser, B. A., Oprea, A., Biggio, B., Pelillo, M., & Roli, F. (2023). Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning. *ACM Computing Surveys*, 3585385. https://doi.org/10.1145/3585385

Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D., & Matheny, M. E. (2017). Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association: JAMIA*, *24*(6), 1052–1061. https://doi.org/10.1093/jamia/ocx030

De Roux, D., Perez, B., Moreno, A., Villamil, M. D. P., & Figueroa, C. (2018). Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 215–222. https://doi.org/10.1145/3219819.3219878

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting Material Accounting Misstatements*. *Contemporary Accounting Research*, *28*(1), 17–82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

Deng, W., & Zheng, L. (2021). *Are Labels Always Necessary for Classifier Accuracy Evaluation?* (arXiv:2007.02915). arXiv. http://arxiv.org/abs/2007.02915

Earley, C. E. (2015). Data analytics in auditing: Opportunities and challenges. *Business Horizons*, *58*(5), 493–500. https://doi.org/10.1016/j.bushor.2015.05.002

El-Hay, T., & Yanover, C. (2022). *Estimating Model Performance on External Samples from Their Limited Statistical Characteristics* (arXiv:2202.13683). arXiv. https://doi.org/10.48550/arXiv.2202.13683

Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N., & Riniker, S. (2021). GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning. *Journal of Chemical Information and Modeling*, *61*(6), 2623–2640. https://doi.org/10.1021/acs.jcim.1c00160

Farooq, Z., Sjödin, H., Semenza, J. C., Tozan, Y., Sewe, M. O., Wallin, J., & Rocklöv, J. (2023). European projections of West Nile virus transmission under climate change scenarios. *One Health*, *16*, 100509. https://doi.org/10.1016/j.onehlt.2023.100509

Ferrell, W. R. (1994). Calibration of sensory and cognitive judgments: A single model for both. *Scandinavian Journal of Psychology*, *35*(4), 297–314. https://doi.org/10.1111/j.1467-9450.1994.tb00955.x

Garg, S., Balakrishnan, S., Lipton, Z. C., Neyshabur, B., & Sedghi, H. (2022). *Leveraging Unlabeled Data to Predict Out-of-Distribution Performance* (arXiv:2201.04234). arXiv. https://doi.org/10.48550/arXiv.2201.04234

Gokcesu, K., & Gokcesu, H. (2021). *Optimally Efficient Sequential Calibration of Binary Classifiers to Minimize Classification Error* (arXiv:2108.08780). arXiv. https://doi.org/10.48550/arXiv.2108.08780

Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., & Schmidt, L. (2021). *Predicting with Confidence on Unseen Distributions* (arXiv:2107.03315). arXiv. https://doi.org/10.48550/arXiv.2107.03315

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). *On Calibration of Modern Neural Networks* (arXiv:1706.04599). arXiv. https://doi.org/10.48550/arXiv.1706.04599

Hsu, Y.-C., Shen, Y., Jin, H., & Kira, Z. (2020, February 26). *Generalized ODIN: Detecting Out-of-distribution Image without Learning from Out-of-distribution Data*. ArXiv.Org. https://arxiv.org/abs/2002.11297v2

Humphrey, A., Kuberski, W., Bialek, J., Perrakis, N., Cools, W., Nuyttens, N., Elakhrass, H., & Cunha, P. A. C. (2022). Machine-learning classification of astronomical sources: Estimating F1-score in the absence of ground truth. *Monthly Notices of the Royal Astronomical Society: Letters*, *517*(1), L116–L120. https://doi.org/10.1093/mnrasl/slac120

Johansson, U., Löfström, T., & Boström, H. (2021). Calibrating multi-class models. *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, 111–130. https://proceedings.mlr.press/v152/johansson21a.html

Joy, T., Pinto, F., Lim, S.-N., Torr, P. H. S., & Dokania, P. K. (2022). *Sample-dependent Adaptive Temperature Scaling for Improved Calibration* (arXiv:2207.06211). arXiv. http://arxiv.org/abs/2207.06211

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Kull, M., Perello-Nieto, M., Kängsepp, M., Filho, T. S., Song, H., & Flach, P. (2019). *Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration* (arXiv:1910.12656). arXiv. https://doi.org/10.48550/arXiv.1910.12656

Kumar, A., Liang, P., & Ma, T. (2020). *Verified Uncertainty Calibration* (arXiv:1909.10155). arXiv. https://doi.org/10.48550/arXiv.1909.10155

Li, G., Wang, T., Chen, Q., Shao, P., Xiong, N., & Vasilakos, A. (2022). A Survey on Particle Swarm Optimization for Association Rule Mining. *Electronics*, *11*(19), Article 19. https://doi.org/10.3390/electronics11193044

Lin, C.-C., Chiu, A., Huang, S. Y., & Yen, D. (2015). Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowledge-Based Systems*, *89*. https://doi.org/10.1016/j.knosys.2015.08.011

Lucena, B. (2018, September 20). *Spline-Based Probability Calibration*. ArXiv.Org. https://arxiv.org/abs/1809.07751v1

Meyer-Bullerdiek, F. (2021). Out-of-sample performance of the Black-Litterman model. *Journal of Finance and Investment Analysis*, 29–51. https://doi.org/10.47260/jfia/1022

Miller, B. A., Vila, J., Kirn, M., & Zipkin, J. R. (2018). Classifier Performance Estimation with Unbalanced, Partially Labeled Data. *Proceedings of The International Workshop on Cost-Sensitive Learning*, 4–16. https://proceedings.mlr.press/v88/miller18a.html

Mosquera, C., Ferrer, L., Milone, D., Luna, D., & Ferrante, E. (2022). *Impact of class imbalance on chest x-ray classifiers: Towards better evaluation practices for discrimination and calibration performance* (arXiv:2112.12843). arXiv. https://doi.org/10.48550/arXiv.2112.12843

NannyML. (n.d.). *Estimation of Performance of the Monitored Model—NannyML 0.8.6 documentation*. Retrieved 23 June 2023, from https://nannyml.readthedocs.io/en/stable/how_it_works/performance_estimation.html#confidence-based-performance-estimation-cbpe

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*, 625–632. https://doi.org/10.1145/1102351.1102430

Nixon, J., Dusenberry, M., Jerfel, G., Nguyen, T., Liu, J., Zhang, L., & Tran, D. (2020). *Measuring Calibration in Deep Learning* (arXiv:1904.01685). arXiv. https://doi.org/10.48550/arXiv.1904.01685

Pasquier, R., & Smith, I. F. C. (2015). Using measurement to reduce model uncertainty for better predictions. In *Structural Engineering: Providing Solutions to Global Challenges* (pp. 1255–1262). https://doi.org/10.2749/222137815818358538

Purda, L., & Skillicorn, D. (2015). Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection. *Contemporary Accounting Research*, *32*(3), 1193–1223. https://doi.org/10.1111/1911-3846.12089

Rabade, S. U. (2022). Use of Machine Learning in Financial Fraud Detection: A Review. *International Journal of Advanced Research in Science, Communication and Technology*, 38–44. https://doi.org/10.48175/IJARSCT-7595

Schelter, S., Rukat, T., & Biessmann, F. (2020). Learning to Validate the Predictions of Black Box Classifiers on Unseen Data. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 1289–1299. https://doi.org/10.1145/3318464.3380604

Song, H., Kull, M., & Flach, P. (2018). *Non-Parametric Calibration of Probabilistic Regression* (arXiv:1806.07690). arXiv. http://arxiv.org/abs/1806.07690

Sugiyama, M. (2016). Introduction to Statistical Machine Learning. In *Introduction to Statistical Machine Learning* (pp. 375–390). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-802121-7.00044-3

Tabacof, P., & Costabello, L. (2020). *Probability Calibration for Knowledge Graph Embedding Models* (arXiv:1912.10000). arXiv. https://doi.org/10.48550/arXiv.1912.10000

Tiittanen, H., Oikarinen, E., Henelius, A., & Puolamäki, K. (2019). *Estimating regression errors without ground truth values* (arXiv:1910.04069). arXiv. http://arxiv.org/abs/1910.04069

Trottini, M., Campus, G., Corridore, D., Cocco, F., Cagetti, M. G., Vigo, M. I., Polimeni, A., & Bossù, M. (2020). Assessing the Predictive Performance of Probabilistic Caries Risk Assessment Models: The Importance of Calibration. *Caries Research*, *54*(3), 258–265. https://doi.org/10.1159/000507276

Uijen, S. (2019). *Unsupervised Error Detection in Accounting Data A Data-driven Audit Approach*.

Vanini, P., Rossi, S., Zvizdic, E., & Domenig, T. (2023). Online payment fraud: From anomaly detection to risk management. *Financial Innovation*, *9*(1), 66. https://doi.org/10.1186/s40854-023-00470-w

Vartak, M. (2021). From ML models to intelligent applications: The rise of MLOps. *Proceedings of the VLDB Endowment*, *14*(13), 3419–3419. https://doi.org/10.14778/3484224.3484240

Xiuguo, W., & Shengyong, D. (2022). An Analysis on Financial Statement Fraud Detection for Chinese Listed Companies Using Deep Learning. *IEEE Access*, *10*, 22516–22532. https://doi.org/10.1109/ACCESS.2022.3153478

Yu, S., Liu, D., Zhang, Y., Zhao, S., & Wang, W. (2021). DPTCN: A novel deep CNN model for short text classification. *Journal of Intelligent & Fuzzy Systems*, *41*(6), 7093–7100. https://doi.org/10.3233/JIFS-210970

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 694–699. https://doi.org/10.1145/775047.775151