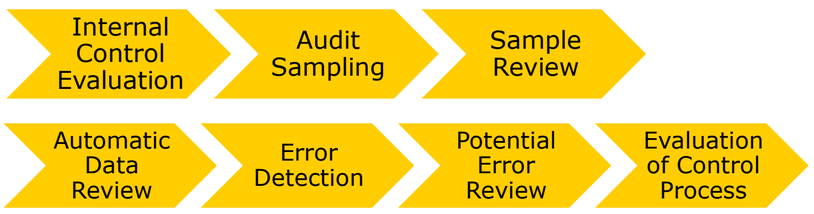
Confidence Based Performance Estimation

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

Alex Essaijan

Introduction

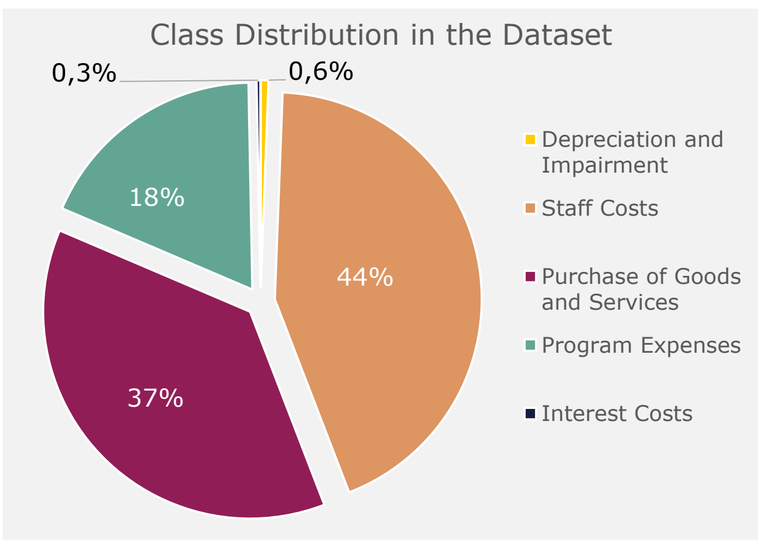
This study aims to estimate the performance of machine learning models on unseen accounting data for financial audits. In real-world scenarios, the assumption that the performance on the test sets represent the production performance may not hold, leading to uncertainty. The CGAS aims to transition from random sampling to a machine learning-informed stratified sampling method for financial audits. This transition would enhance audit efficiency and effectiveness by leveraging the model's capability to examine every transaction.



Research Question: *How can the performance of machine learning models be estimated on unseen accounting data?*

Data

The dataset consists of the transactions of two Dutch ministries during in 2017 & 2018 and was collected by the Central Government Accounting Service. It was divided into three subsets: the training set, calibration set and test set, with a random split of 70:15:15. The complete dataset has a total of 761,694 transactions. Divided over 5 distinct classes. Note: The class label distributions in each subset are highly comparable.



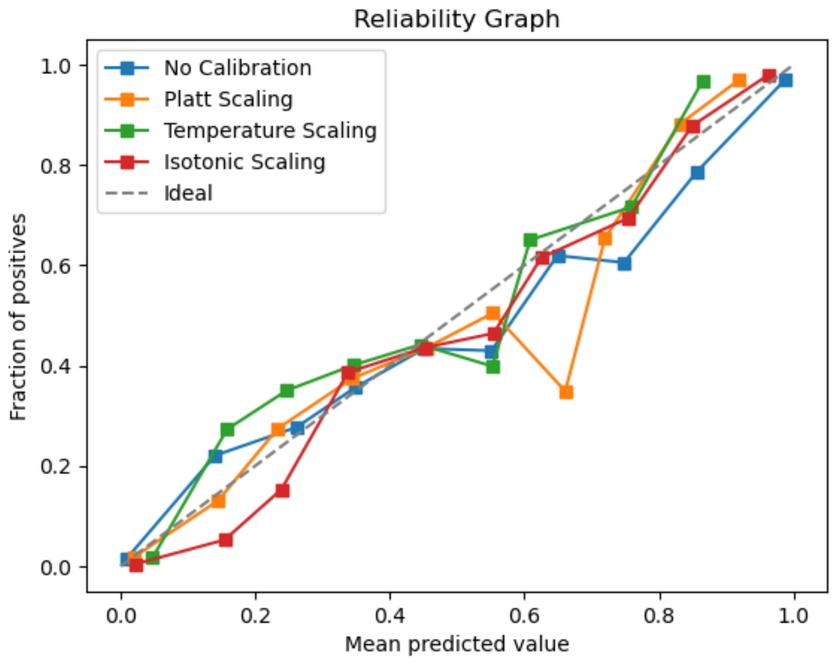
Method

A CNN model was trained to classify transaction descriptions into specific classes. The probabilities generated by the classifier were then calibrated using three different methods to improve their accuracy. The best calibrated probabilities were selected based on the lowest values of Log Loss, ECE, and Brier score.

This performance was estimated with the Confidence-Based Performance Estimation (CBPE) approach. Specific performance metrics were calculated using the predicted labels and probabilities, including the False Positive Rate (FPR). FPR was determined by identifying instances classified as positive but below a threshold (0.62) and dividing the False Positives by the sum of False Positives and True Negatives for each class.

Results

The research aimed to investigate the estimation of a confusion matrix in multiclass classification but faced challenges due to the absence of ground truth labels. Another method to estimate the performance of a classifier without truth labeled is the CBPE method. One of its assumptions is that the probabilities used as input are well calibrated. For this reason, probability calibration was assessed using different methods and performance metrics. As the uncalibrated probabilities demonstrated the best calibration performance on the performance metrics, they were selected for the CBPE method.



FPR vs Estimated FPR per class

Class label	FPR	Estimated FPR
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

Comparing the estimated FPR with the truth values using CBPE, the estimated values stayed within 2% deviation. However the Program Expenses class has a significant mismatch, with a 19.95% ground truth FPR and a 3.66% estimated FPR.

Calibration Methods Compared on Performance Metrics

Method	Log Loss	ECE	Mean Brier Score	Accuracy
Uncalibrated	0.486	0.092	0.155	0.810
Platt Scaling	0.535	0.124	0.141	0.808
Temperature Scaling	0.598	0.121	0.122	0.810
Isotonic Scaling	0.544	0.103	0.150	0.810

The estimated classification report mostly shows higher estimated scores than true values across classes, indicating an optimistic estimation of the model's performance in class identification. However, these estimations still show a reasonable level of similarity. While some outliers can be observed in the table below.

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

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

Discussion

The study found that the uncalibrated model performed best in terms of reflecting the true probabilities. This suggests that the original probability estimates were already well-calibrated. The CBPE method consistently provided estimations with an error range of approximately 3% compared to the true performance. However, discrepancies were found for the class Program Expenses, causing higher deviations in recall and FPR. These higher deviations can be explained by the model's performance issues for this specific class. These findings have practical implications for auditors, as they can estimate machine learning model performance without ground truth, allowing them to consider of integrating machine learning into auditing.

Limitations & Future Research

This paper has limitations due to bias, time and scope constraints. The reliability of the model is determined by the underlying classifier's performance and the data checked by auditors, which may be susceptible to human error. The CBPE method's binary approach may oversimplify complexities and interdependencies between classes. Future research could explore methods that extend beyond binary transformation for multiclass classification and could address class imbalance. The high false positive rate, particularly for Program Expenses, negatively impacts estimations, needing further investigation. Future research could focus on real-time monitoring of model performance, addressing data, concept drift challenges and improving robustness in real-world applications.