Quality & Quantity
https://doi.org/10.1007/s11135-026-02725-x

Risk-based predictive modelling for audit verification: evidence from EU-funded programmes

Elisa Verna[1] · **Gianfranco Genta**[1] · **Maurizio Galetto**[1]

Received: 1 December 2025 / Accepted: 22 March 2026

Abstract
This study proposes a machine learning framework to support risk-based verification of expenditure declarations in European Structural and Investment Funds, reflecting the current regulatory emphasis on proportional and data-driven audit strategies. Quantitatively, the problem is formulated as an imbalanced three-class classification task with ordered outcomes on high-dimensional administrative data; the ordinal structure is exploited ex post in evaluation and error interpretation. The framework classifies expense documents as validated, partially validated, or not validated, and provides audit authorities with interpretable probability estimates for each case. A predictive model was trained and validated on more than ninety thousand expense documents from the Italian Regional Operational Programme co-funded by the European Regional Development Fund (2014–2020). Methodological challenges—ordered outcomes, severe class imbalance, and mixed-type features—were addressed through targeted preprocessing and the CatBoost gradient-boosting algorithm. The model achieved satisfactory predictive performance, offering probabilistic outputs aligned with the ordered structure of audit outcomes. Variable-importance analysis confirmed the relevance of both financial and administrative variables in predicting irregularities. The framework is designed with operational integration in mind and could underpin risk-based sampling in expenditure verification, subject to further validation across time, programmes, and beneficiary structures. Departing from a literature that largely focuses on binary classification or fraud detection, the study addresses the understudied challenge of multi-class prediction in public expenditure control and provides an interpretable prototype decision-support tool. The model could support public authorities in prioritizing controls and allocating resources more efficiently, contributing to the modernization of European Union fund management and promoting data-driven, proportionate oversight—conditional on governance arrangements and external validation.

Keywords Risk-based verification · Predictive analytics · Machine learning · Public sector performance · Data-driven decision-making · EU fund management

Extended author information available on the last page of the article

Published online: 28 March 2026

1 Introduction

The 2021–2027 programming period for European Structural and Investment Funds (ESIF) introduces a significant shift in how expenditure controls are conducted. Article 74(2) of the Common Provisions Regulation (EU) 2021/1060 mandates that management verifications be risk-based and proportionate to the level of risk identified ex ante (European Commission 2023). This marks a move away from exhaustive 100% checks toward a more proportionate, risk-focused allocation of control resources.

From a methodological perspective, risk-based verification can be formalised as an imbalanced three-class prediction problem with ordered outcomes and mixed-type administrative data, a configuration that remains underexplored in applied quantitative research. Recent contributions employing ordinal or priority-based quantitative methods in applied contexts further highlight the relevance of structured ordinal modelling frameworks for complex decision processes (Nawaz et al. 2025); however, the predictive engine adopted in this paper is a standard multiclass classifier and the ordinality is handled ex post. This methodological gap mirrors the practical challenges faced by managing authorities, who lack operational tools capable of translating such data patterns into reliable ex-ante risk estimates. As a result, the adoption of risk-based verification remains limited, particularly at regional and local levels, and authorities struggle to identify which declarations are more likely to contain irregularities. Developing predictive capabilities is therefore essential for enabling a transition toward proportionate, data-driven controls. Without such tools, verification efforts remain inefficient and risk-averse, undermining the intent of the regulatory shift. This implementation and methodological gap motivate the present study.

This study proposes a machine-learning classification framework to support risk-based expenditure verification. The model estimates the probability that each expense document will be fully validated, partially validated, or rejected, allowing managing authorities to prioritise high-risk items while streamlining checks for low-risk ones.

The study pursues two complementary objectives. The predictive objective is to estimate tri-class verification outcomes under severe class imbalance using administrative microdata. The decision-support objective is to illustrate how predicted risk can inform selective verification policies. The former is directly validated on held-out data, whereas the latter is illustrated through scenario-based analyses and remains conditional on implementation choices and further external validation.

The predictive engine is implemented as a standard three-class classifier; the ordered nature of audit outcomes is incorporated ex post through evaluation and error-severity reporting (e.g., one-step vs. two-step misclassifications) and through probability-based decision rules.

The framework is validated through a real-world case study using administrative data from a regional Operational Programme in Italy. The dataset, drawn from the 2014–2020 European Regional Development Fund (ERDF) program, includes tens of thousands of expense documents with known verification outcomes. Although the model is trained on past-period data, the study is motivated by the 2021–2027 regulatory context, where risk-based control is mandatory. The case study illustrates the potential of the proposed risk scoring approach as a decision-support component, while keeping any operational integration explicitly conditional on governance choices and validation across programmes and time.

The contributions of this work are both methodological and practical. Methodologically, the framework addresses a combination of real-world challenges often overlooked – namely, ordered outcome classes, severe class imbalance (few irregular cases among many valid ones), and heterogeneous, high-dimensional data (including many categorical fields). Prior research in public fund auditing and fraud detection has often simplified the problem, treating it as binary classification on balanced datasets—assumptions that do not hold here (Aldana et al. 2022; Siciliani et al. 2023). A similar trend exists in predictive modelling in manufacturing and quality control, where models are typically designed for continuous outputs or binary decisions under more controlled conditions (Galetto et al. 2020; Verna et al. 2021, 2023). By contrast, the present framework jointly addresses ordered outcome classes, skewed class distributions, and mixed data types, reflecting the specific constraints of administrative verification. Practically, the model provides an empirical basis for prioritising verifications in resource-constrained environments.

More broadly, the work aligns with the literature emphasising the need for improved performance measurement and data-driven decision-making in the public sector (Garengo and Sardi 2021; Radnor and McGuire 2004). Recent literature also highlights the institutional and accountability constraints that shape the adoption of analytics in public organisations (Ensslin et al. 2022; Halachmi and Greiling 2011). In this context, Machine Learning (ML) can support more structured and traceable prioritisation of verification activities, in line with research on how performance appraisal mechanisms influence auditors' behaviour and effectiveness (Henriquez-Machado et al. 2025; Nehme et al. 2020).

To operationalise this perspective, the study adopts modern ML techniques suited to mixed-type administrative data. Conventional approaches (e.g., logistic regression or ordinal logistic models) assume linearity, balanced classes, and mostly numeric inputs—conditions rarely met in this setting; accordingly, preliminary Ordered Logistic Regression experiments showed poor performance (low pseudo-R^2 and weak minority-class detection). In contrast, the proposed model uses CatBoost (Prokhorenkova et al. 2017), which natively handles high-cardinality categorical variables and mitigates overfitting in gradient boosting. Combined with targeted preprocessing for missingness and data-quality screening, the approach yields a robust and interpretable predictor, providing class probabilities that can support risk-based audit planning.

The remainder of the paper is structured as follows. Section 2 situates the study within the literature on risk-based control and predictive analytics in public finance. Section 3 details the dataset, preprocessing pipeline and modelling choices. Section 4 reports predictive results and scenario analyses. Section 5 discusses implications, limitations and future work, and Sect. 6 concludes the paper.

2 Background and literature review

2.1 Regulatory context: risk-based verification

The 2021–2027 programming period for EU Structural Funds introduced a regulatory shift mandating risk-based management verifications, as per Article 74(2) of Regulation (EU) 2021/1060. This marks a move from exhaustive checks to proportionate controls based on ex ante risk assessments. Authorities are expected to intensify checks on high-risk opera-

tions and reduce oversight on low-risk ones, balancing budget protection with administrative efficiency.

In comparison to the 2014–2020 programming period (when such a risk-driven approach was encouraged but not strictly required), the 2021–2027 period makes risk-based verification a legal obligation. The European Commission's guidance emphasizes that this approach should significantly reduce the administrative burden for compliant beneficiaries (since fewer controls will be needed for low-risk cases) while improving the effectiveness of control systems by targeting problematic areas instead of expending time on every single expense document (European Commission 2023). In principle, this should allow Managing Authorities (MAs) to allocate their audit and verification efforts more judiciously, focusing on quality of checks rather than sheer quantity.

Nonetheless, implementation has been uneven, especially at subnational levels. Many authorities continue to perform exhaustive or fixed-percentage verifications, lacking tools to assess risks upfront. Factors such as beneficiary profile, project type, or funding amount complicate predictions. The absence of reliable risk models often leads to hesitation in abandoning full-scope checks, undermining the intent of the regulation.

Predictive analytics can fill this gap by enabling early detection of potentially irregular expense documents. Tools like ARACHNE, developed by the European Commission (2024), exemplify this approach. However, standardized tools may overlook regional specificities or lack transparency. This motivates the development of context-specific risk-scoring models on administrative micro-data, whose outputs can be embedded into operational decision rules while remaining auditable and interpretable. There is thus scope for context-specific, interpretable models to support proportionate and effective verification in line with Article 74(2) (European Commission 2023).

2.2 Predictive analytics in public financial management

Within public financial management, the strand most relevant to EU management verifications concerns ex ante risk scoring on administrative micro-data to prioritise controls under limited audit capacity. This evidence base shows that supervised models can detect patterns consistent with irregular or anomalous cases, but most applications remain framed as binary detection (fraud/no-fraud) rather than modelling the ordered multi-class outcomes typical of expenditure verification.

(i) Risk scoring on administrative audit data. Case studies in procurement and contracting show that machine-learning models trained on administrative records can flag higher-risk transactions and complement human screening. Bai and Qiu (2023) and Aldana et al. (2022), for example, illustrate how non-linear learners can exploit complex interactions and relational descriptors to improve detection performance. While informative for feasibility, these contributions typically abstract away from the operational reality of verification, where outcomes may include full validation, partial validation, or rejection rather than a single fraud label.

(ii) Decision-support evaluation of risk models. A second, closely related stream emphasises that predictive accuracy alone is insufficient when model outputs are used to allocate scarce verification effort. Studies proposing dashboard-based or triage-oriented tools (e.g., Siciliani et al. (2023) highlight that operational value depends on explicit

decision rules linking predicted risk to audit intensity, together with transparent reporting of the workload–residual-risk trade-off. Accordingly, evaluation should be reported not only as classification performance, but also as policy-relevant quantities under explicit assumptions (e.g., audit budgets and error severity), and—when probabilities are used—through calibration diagnostics and proper scoring rules (e.g., reliability analysis and Brier-type scores).

(iii) Hierarchical dependence and leakage. Administrative audit datasets are typically clustered (multiple documents per beneficiary and project), which can inflate performance if the validation split allows correlated records to appear across training and test partitions. This motivates group-aware validation designs that respect the data-generating structure and avoid optimistic estimates due to within-entity dependence (Guignard et al. 2024; Kapoor and Narayanan 2023).

(iv) Concept drift and temporal validation. In regulated environments, risk patterns may change over time due to evolving rules, auditing practices, beneficiary behaviour, or documentation quality. This form of temporal drift (concept drift) motivates time-aware validation (e.g., training on earlier cohorts and testing on later ones) when models are intended for prospective use (Gama et al. 2014a).

Against this background, the gap addressed here is the development and evaluation of an interpretable tri-class risk-scoring model on imbalanced administrative micro-data, explicitly connected to decision-support evaluation under audit-capacity constraints, rather than to fraud/no-fraud detection alone.

2.3 Methodological challenges in EU expenditure verification

Applying predictive analytics to EU expenditure verifications presents distinct methodological challenges that go beyond standard ML settings. Four main issues are particularly relevant:

(1) Heterogeneous and high-cardinality variables. Fund management datasets often include a large number of categorical variables, such as project type, region, or funding instrument, many with high cardinality. This heterogeneity complicates preprocessing and increases the risk of overfitting, especially when using simplistic encoding techniques. As highlighted by Liang (2025), modern encoding strategies like target encoding or embeddings are more effective in preserving information and enhancing model interpretability. Hence, model selection must prioritize algorithms capable of natively handling such data.

(2) Ordinal target variable. Audit outcomes in EU verifications are not nominal but inherently ordinal: “validated” is better than “partially validated”, which is better than “not validated.” However, many classification models ignore this structure, leading to misclassifications with unequal severity. As noted by Lázaro and Figueiras-Vidal (2023), ordinal classification under imbalance remains a challenging task. Bonnier et al. (2022) further stress that decomposing ordinal problems into binary ones leads to suboptimal results. In this study, the predictive engine is implemented as a three-class probabilistic classifier. The ordered nature of the outcomes is accounted for at the evaluation and

interpretation stage, by distinguishing misclassifications by their ordinal distance and by informing the design of decision rules.

(3) Severe class imbalance. In compliance audits, most cases are regular, while irregularities, especially full rejections, are rare. This imbalance biases standard classifiers toward the majority class, reducing sensitivity to the high-risk minority. Lemnaru and Potolea (2012) emphasize that no universal solution exists; appropriate handling may involve resampling, cost-sensitive learning, or robust ensemble models. In the current study, resampling strategies are evaluated to ensure minority classes remain detectable without inflating false positives.

(4) Missing and noisy data. Administrative datasets often suffer from missing values and inconsistencies due to fragmented data collection processes. While some ML models tolerate limited missingness, extensive gaps require proper imputation to avoid bias. Wongkamthong and Akande (2020) demonstrate that model-based approaches, such as multiple imputation using chained equations, perform better than simplistic methods in preserving data integrity, particularly with ordinal features. In this study, dedicated preprocessing steps were introduced to manage missingness without distorting variable relationships.

These challenges define the requirements for a reliable risk-prediction system in EU expenditure control. The proposed framework addresses them by combining imputation, modelling choices suited to mixed-type administrative data, and imbalance-handling strategies tailored to tri-class verification outcomes. Importantly, administrative audit data are often hierarchically structured (e.g., multiple documents per beneficiary and project), which can create dependence across observations and requires validation designs that avoid information leakage across splits. Consistently with this, the empirical validation in this study adopts a group-aware split at beneficiary level to mitigate within-entity dependence (Guignard et al. 2024; Kapoor and Narayanan 2023). Moreover, because administrative processes may exhibit temporal drift, time-aware validation is also relevant when the intended operational context differs from the training period (Gama et al. 2014a).

Against this background, the specific gap addressed by this study is the development and evaluation of an interpretable tri-class risk-scoring model on imbalanced administrative micro-data, explicitly linked to decision-support evaluation under audit-budget constraints (workload–risk trade-offs), rather than to fraud/no-fraud detection alone.

3 Methodology

3.1 Dataset and problem definition

The analysis is based on a comprehensive administrative dataset of $N = 91{,}924$ expense documents from the Piedmont Region's POR FESR 2014–2020 (European Regional Development Fund program). Each record corresponds to a single expense document (a claim for reimbursement of project expenses) and is associated with an audit outcome. The data hierarchy is such that multiple expense documents compose a declaration of expenditure of a beneficiary, and multiple declarations relate to a project. For this study, the unit of analysis is the individual expense document, as this is the level at which validation outcomes are

recorded. The declaration dates span 2017–2023, reflecting the multi-year implementation and closure of the 2014–2020 programme.

Each document has a categorical audit outcome label (denoted *outcome_exp_doc*), which serves as the target variable. The possible outcomes are defined as follows:

- VALIDATED: the expense document was fully accepted (no irregularities).
- PARTIALLY VALIDATED: the expense document was reviewed and some expenditures were disallowed (only part of the amount was deemed eligible).
- NOT VALIDATED: the expense document was entirely rejected or found ineligible.

These outcomes have an inherent ordinal relationship (NOT VALIDATED < PARTIALLY VALIDATED < VALIDATED in terms of favorability). However, the prediction task is framed as a three-class classification problem. The ordinal nature is acknowledged in evaluation (treating some misclassification errors as more severe than others), but no strict ordinal regression algorithm is imposed. The objective is to estimate, for each new expense document, the probabilities of belonging to each outcome class. This enables a risk score for each document—for example, the predicted probability of an adverse outcome (partial or non-validation). Audit managers can use these risk scores to prioritize documents for manual verification, implementing a risk-based selection strategy. In practice, documents with a high predicted risk of irregularities would be audited with higher priority, while those predicted to be low-risk could undergo minimal or sample-based checks.

The methodological workflow unfolds in four main steps: (i) data preparation and preprocessing; (ii) feature engineering and selection of predictive variables; (iii) model training with alternative imbalance-handling strategies; and (iv) evaluation and scenario analysis, as detailed in Sects. 3.2–3.7.

3.2 Data preparation and preprocessing

A structured data preparation pipeline was followed to ensure data quality and to handle complexities in the dataset. Key steps included:

- Missing Value Imputation: Rather than dropping records with missing fields (which could bias the analysis or reduce the already small minority classes), an iterative multivariate imputation strategy was applied. Specifically, the *Iterative Imputer* from scikit-learn was used (Pedregosa et al. 2012), implementing a chained equations approach (MICE). This method initializes missing values (e.g. with mean or mode) and then iteratively refines those estimates by training predictive models for each incomplete feature in turn, using all other features as predictors. Iterative imputation helps preserve inter-variable relationships better than simple mean imputation and is well-suited for datasets with a mix of categorical (dummy-coded) and continuous variables (van Buuren 2018). To prevent leakage, imputation models were fitted within the training data only (i.e., within each cross-validation fold during model selection, and then refit on the full training split before transforming the held-out test set).
- Outlier screening: Extreme values in key monetary fields may stem from legitimate high-value expenditures or from data-entry/transcription issues. In this setting, Grubbs' test (Hodge and Austin 2004) was used only as a flagging mechanism to produce a short

list of cases for targeted human plausibility checks against the supporting documentation; flagged records were neither capped nor removed. In the analysed dataset, all flagged high-value observations were verified as legitimate, so no values were modified and the modelling sample remained unchanged.

- Exploratory checks were conducted to understand data quality and guide modelling choices (e.g., missingness patterns, plausibility of monetary fields, and the presence of dependence across beneficiaries/projects).
- Feature Encoding: Many variables were categorical (nominal variables like "legal form of beneficiary" or "province"). Feature encoding refers to the process of transforming such categorical variables into a numerical format suitable for model learning (Zheng and Casari 2018). Rather than applying traditional encoding techniques like one-hot encoding, which can lead to high dimensionality and sparsity, this study leveraged CatBoost's native encoding mechanism. CatBoost performs ordered target statistics encoding, where categorical values are replaced by conditional statistics derived from the data in a way that avoids target leakage and preserves predictive signal (Prokhorenkova et al. 2017). This allowed the inclusion of a large number of categorical variables without extensive preprocessing. For a few high-cardinality variables (e.g., a free-text field or ID that appears very few times), additional consideration was given to exclusion or transformation, but in practice, CatBoost handled all selected categorical variables effectively. This capability strongly influenced variable selection, allowing the retention of features that would otherwise have been dropped or heavily transformed using alternative algorithms.

After these preprocessing steps, a cleaned and enriched dataset was obtained for modeling. No feature scaling was necessary for the tree-based models (which are scale-invariant), while standardisation was applied only for the logistic-regression benchmarks. The dataset was then split into training and test partitions using an 80/20 beneficiary-grouped hold-out design (80213 training documents; 11711 test documents). All documents associated with the same beneficiary were assigned to a single partition to prevent information leakage due to within-beneficiary dependence (Kaufman et al. 2012). The beneficiary identifier was used exclusively for splitting and was never included as a predictor. Model selection within the training partition relied on StratifiedGroupKFold cross-validation ($n=3$), which preserves beneficiary separation while approximately maintaining the class distribution across folds; all preprocessing and resampling steps were executed within each training fold through a pipeline.

3.3 Feature engineering

Through data collection and domain consultation, a set of 21 predictive variables was identified based on availability, quality, and relevance. These features capture various aspects of each expense document, including the nature of the expense, financial amounts, project attributes, and beneficiary characteristics. Table 1 provides an overview of the variables used in the model, along with a brief description and data type (quantitative or categorical). Purely identifying information (IDs, names of projects or beneficiaries) and any variables not available at the time of verification planning were excluded from modeling. Notably, the field indicating the actually validated amount on the document was omitted to prevent data

Table 1 Predictive variables used in the model

Variable name	Description	Type
doc_type	Type of expense document	Categorical
reported_amount_doc	Amount actually reported on the expense document	Quantitative
budgeted_amount_item	Amount allocated in the project budget for that type of expenditure item	Quantitative
expense_category	Expense category (broad category of cost)	Categorical
aid_type	Type of aid (contribution or funding)	Categorical
activity_days	Days of activity of the beneficiary	Quantitative
legal_form	Legal form of the beneficiary	Categorical
eligible_amount_project	Total amount of the project, on which the facilitation granted is calculated	Quantitative
granted_aid_amount	Amount of actual aid: equals the disbursed amount for funding, or a percentage of the eligible amount for contributions	Quantitative
project_duration	Project duration (months)	Quantitative
reported_amount_declaration	Amount actually reported on the declaration of expenditure	Quantitative
aid_intensity_pct	Percentage of aid to the project (aid intensity)	Quantitative
declaration_type	Type of declaration of expenditure (intermediate, final, supplementary)	Categorical
enterprise_size	Enterprise size (SME, large, etc.)	Categorical
call_type	Call type (investment, research and development, etc.)	Categorical
consultant_flag	Consultant involved (binary flag if an external consultant assisted in the project)	Categorical
macro_sector	Macro-sector of the beneficiary's activity	Categorical
province	Province of the beneficiary	Categorical
beneficiary_risk_score	Risk score based on beneficiary characteristics, computed using the Arachne risk-scoring tool for EU fund management (European Commission 2024)	Quantitative
partnership_flag	Partnership flag (whether the project involves multiple partners)	Categorical
advance_flag	Advance payment flag (whether advance funds were given)	Categorical

leakage, as it directly reflects the outcome (including it would allow the model to trivially "predict" the result using a post-outcome variable).

These features collectively provide a broad representation of each case: from the nature of the expense (doc_type, expense_category) to key financial figures (reported_amount_doc, budgeted_amount_item, eligible_amount_project, etc.), to project data (call_type, project_duration), and beneficiary attributes (legal_form, enterprise_size, province, etc.).

3.4 Outcome distribution and class imbalance

As anticipated in a public audit context, the class distribution of outcomes is highly imbalanced. Table 2 summarizes the frequency of each audit outcome in the dataset. The vast majority of expense documents (almost 87%) were fully validated, whereas under 4% were not validated. Partially validated cases constitute the remainder (about 9.5%).

This extreme skew is a significant challenge: a naïve classifier that always predicts "VALIDATED" would achieve about 86.6% overall accuracy, yet would completely fail to

Table 2 Class distribution of audit outcomes in the dataset (N=91924 documents)

Audit outcome	Frequency	Percentage (%)
VALIDATED	79,584	86.58
PARTIALLY VALIDATED	8729	9.50
NOT VALIDATED	3611	3.93

detect the minority classes of interest (catching 0% of the problematic cases). Addressing the imbalance was therefore critical. A two-pronged approach was taken: firstly, choosing a modeling algorithm known to be robust to class imbalance (CatBoost, as introduced later); and secondly, applying explicit resampling strategies during training (Sect. 3.6). These strategies ensured that the model would not simply be overwhelmed by the dominant class.

Given the imbalance and ordinal nature of the target, an initial decision was whether to treat the task as a proper ordinal regression or as a multi-class classification. While specialized ordinal classification methods exist (e.g. ordinal logistic regression, ordinal decision trees), it was found that a well-tuned multi-class approach can perform equivalently for this application. The ordinal structure was instead incorporated at the evaluation stage, recognizing, for example, that mistaking a NOT VALIDATED for a PARTIALLY VALIDATED is less severe than mistaking it for a fully VALIDATED. For simplicity, the model was developed as a multi-class classifier, with the understanding that the results would be interpreted with the outcome ordering in mind.

3.5 Model selection and training

The search for an effective classifier followed a two-stage path. In the exploratory stage, eight algorithms were benchmarked to capture the full spectrum of methods normally considered in audit analytics: a multinomial logistic regression and its ordinal logit variant as linear baselines; a single CART decision tree to obtain an easily interpretable non-linear reference; a Random Forest and an XGBoost ensemble to represent mainstream tree-based learners; and two CatBoost configurations, combined respectively with oversampling and random undersampling to assess imbalance-handling trade-offs; finally, a majority-class baseline served as a sanity check under severe imbalance.

Each algorithm was trained and evaluated under a leakage-safe design. A beneficiary-disjoint split was used to define the training partition, leaving an untouched beneficiary-disjoint hold-out set for final testing. During model selection, performance was estimated by 3-fold StratifiedGroupKFold cross-validation with beneficiary grouping within the training partition, thereby avoiding leakage while mitigating distributional shifts induced by extreme imbalance. All transformations (encoding and standardisation for linear baselines; resampling for imbalanced learning) were embedded in pipelines fitted on the training folds only. Tree-based learners required minimal preprocessing; CatBoost, in particular, ingested the raw categorical fields and applied its internal encoding scheme suited to high-cardinality administrative variables.

Oversampling and undersampling were applied strictly within the cross-validation folds to mitigate class imbalance (see Sect. 3.6 for details). Hyper-parameter optimisation relied on RandomisedSearchCV (Prokhorenkova et al. 2017). For CatBoost, the search space included the number of boosting iterations, depth, learning rate, L2 regularisation and random strength. A governance-oriented scoring function was adopted to align tuning with the verification objective: macro-average recall was penalised by the severe error rate (NOT

VALIDATED → VALIDATED), thereby discouraging configurations that would auto-validate highly irregular documents. After tuning, the selected configuration was refit on the full training partition under the chosen resampling strategy.

A separate beneficiary-disjoint calibration subset was used to tune the probability threshold governing auto-validation (Sect. 4.3), while preserving the original class proportions for threshold selection. In addition, post-hoc probability calibration (temperature scaling) was fitted on the same beneficiary-disjoint calibration subset and applied to probability outputs used for threshold-based policies (Sect. 4.4). As a robustness check, an ablation experiment re-estimated the CatBoost model without beneficiary_risk_score, keeping the leakage-safe validation protocol unchanged, to quantify its marginal value and address potential circularity.

3.6 Handling class imbalance with oversampling and undersampling

Class imbalance was addressed through two alternative resampling techniques (More 2016) applied only to the training folds:

- Oversampling the minority classes using SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous variables). SMOTENC is an adaptation of SMOTE that can handle mixed data types, generating synthetic examples for minority classes by interpolating between existing instances (Chawla et al. 2011) while respecting categorical variables. SMOTENC was used to generate synthetic examples for the PARTIALLY and NOT VALIDATED classes.
- Undersampling the majority class (VALIDATED) by randomly removing a significant portion of its instances, to balance the class distribution more closely.

Each resampled data set served to train a dedicated CatBoost model. The comparative effectiveness of the two approaches is reported in Sect. 4.

3.7 Evaluation metrics and interpretability

To evaluate the model's performance in a meaningful way, metrics were chosen that account for class imbalance and the ordinal aspect of the outcomes (Rainio et al. 2024). Key evaluation measures included:

- Weighted Accuracy: The overall accuracy, weighted by class frequencies. With such imbalance, accuracy is interpreted with caution, since a high accuracy can be achieved by trivially predicting only the majority class. Weighted accuracy is therefore reported primarily for completeness, while the main emphasis is placed on recall-oriented metrics and the reduction of critical false negatives.
- Precision, Recall, F1-Score (Weighted Average): For each class, precision and recall were computed, then a weighted average (by class support) was taken. The weighted precision/recall/F1 thus reflect performance across all classes in proportion to their prevalence. This approach ensures that the large VALIDATED class is appropriately weighted in the metric, while additional attention is given to the performance on minority classes.

- Macro-Averaged Precision/Recall: In addition to weighted averages, macro-average metrics (unweighted mean across the three classes) were considered, especially macro Recall. This treats each class equally regardless of size, providing insight into how well the model balances the three outcomes. Macro recall was used as a tuning objective for model selection, as mentioned.
- Confusion Matrix: Confusion matrices were examined to see the distribution of predictions versus true outcomes. This gives a granular view of the types of errors made. For instance, it shows how often NOT VALIDATED documents were misclassified as PARTIALLY VALIDATED or as VALIDATED, etc. The confusion matrix is especially useful for understanding whether errors tend to be *adjacent* in the ordinal scale or truly out-of-order.
- Severe error rate (NOT VALIDATED → VALIDATED). Given the ordered outcome and the audit governance context, a severe misclassification is defined as predicting VALIDATED for a document that is actually NOT VALIDATED. The severe NOT→VALID rate is computed as the share of such errors over all evaluated documents; in reporting, the corresponding class-conditional rate (errors divided by the number of NOT VALIDATED documents) is also provided where policy interpretation requires it.

Beyond pure performance, interpretability is crucial for adoption by public authorities. This was addressed by providing:

- Class Probability Outputs: CatBoost returns, for each document, a probability vector over the three outcomes. These probabilities can be used (i) to rank documents by risk and (ii) to implement threshold-based triage rules. Because such policies depend on the probability scale, calibration is assessed on held-out data using reliability diagrams, the multiclass Brier score (a proper scoring rule), and the Expected Calibration Error (ECE), defined as the frequency-weighted average gap between predicted confidence and empirical accuracy across probability bins (Sect. 4.4) (Guo et al. 2017; Murphy and Winkler 1992). For reporting and governance, uncertainty is summarised at the aggregate level through standard confidence intervals on key metrics (i.e., Wilson intervals for precision and recall), not as per-document prediction intervals in operational use.
- Variable Importance: Variable importance was extracted from the CatBoost model (based on its built-in calculation of how much each variable contributes to reducing loss). This is presented as a ranking of the most influential variables in the model's decisions. It helps validate that the model is using reasonable criteria. The model's top variables were found to be intuitive: financial variables (like amounts and scores) and certain project characteristics had the highest importance, aligning with expert expectations.
- Case Studies (Error Analysis): A few cases of *false negatives* (documents that were not validated in reality but the model predicted as validated) were examined in greater detail to understand why the model missed them. Examining these cases with domain experts can sometimes reveal either quirks in the data or areas for improvement (such as needing an additional variable to capture a specific scenario).

3.8 Temporal validation and drift assessment

Because the model is trained on 2014–2020 programme data (with declaration dates spanning 2017–2023) but intended to support verification under the 2021–2027 regulatory context, the beneficiary-grouped split is complemented with a time-based evaluation aimed at detecting temporal drift (changes in rules, auditing practice, beneficiary behaviour, or documentation quality) (Gama et al. 2014b). To avoid reintroducing within-beneficiary leakage in the temporal evaluation (i.e., beneficiaries appearing in both training and test across years), the rolling-origin test set for year t is restricted to "new beneficiaries only", defined as beneficiaries whose first appearance in the dataset occurs in year t; all observations from these beneficiaries are excluded from the training window (years $< t$). All preprocessing and resampling steps are fitted on the training period only. Temporal results and calibration diagnostics are reported in Sect. 4.4.

Section 4 presents the key results for the final CatBoost model (trained with the undersampled data) and illustrates how the resulting risk scores can be mapped to verification strategies under explicit governance assumptions (audit capacity, decision thresholds, and monitoring).

4 Results

4.1 Benchmark models and cross-validation performance

Table 3 reports benchmark performance for baseline algorithms under the same leakage-safe, beneficiary-grouped cross-validation protocol used for the primary classifier (StratifiedGroupKFold within the training FIT partition). This ensures comparability across models when observations are clustered by beneficiary. In the table, severe NOT→VALID rate denotes the proportion of observations with actual NOT VALIDATED that are predicted as VALIDATED (i.e., false auto-validations of fully rejected documents), computed over all documents in the evaluation fold.

While linear baselines and majority prediction achieve high apparent accuracy under severe imbalance, their performance is largely driven by the dominant VALIDATED class

Table 3 Group-aware cross-validation benchmarks for candidate algorithms (training partition; StratifiedGroupKFold with beneficiary grouping)

Algorithm (group-aware CV)	Accuracy	Weighted F1	Macro recall	Severe NOT→VALID rate
CatBoost + undersampling	0.5777	0.6484	0.5509	0.0148
CatBoost + SMOTENC (oversampling)	0.6878	0.7318	0.4956	0.0224
Decision tree (CART)	0.7707	0.7798	0.4426	0.0270
XGBoost	0.8725	0.8332	0.4303	0.0303
Random forest	0.8730	0.8332	0.4251	0.0314
Multinomial logistic regression	0.8602	0.8082	0.3505	0.0382
Ordinal logistic regression (LogisticIT)	0.8645	0.8070	0.3472	0.0382
Majority baseline (always VALIDATED)	0.8668	0.8051	0.3333	0.0399

and yields limited minority sensitivity, as reflected in macro recall close to the majority baseline. Tree-based methods improve minority detection to varying degrees, but their operating characteristics differ substantially when evaluated under leakage-safe beneficiary grouping. In particular, CatBoost with random undersampling achieves the best governance-aligned profile, combining the highest macro recall (0.551) with the lowest severe NOT VALIDATED → VALIDATED error rate (0.0148) among the tested methods. Oversampling via SMOTENC increases overall accuracy relative to undersampling but yields a lower macro recall (0.496) and a higher severe error rate (0.0224). A single CART tree is interpretable but performs less robustly (macro recall 0.443) and exhibits a higher severe error rate than the CatBoost variants. Finally, despite strong weighted F1 and accuracy, XGBoost and Random Forest exhibit substantially higher severe error rates (0.030–0.031) and lower macro recall (0.425–0.430) under group-aware evaluation. These benchmarks motivate selecting the CatBoost + undersampling configuration for the final leakage-safe pipeline, with subsequent threshold calibration and decision-support back-testing reported in Sects. 4.3–4.4. Because the evaluation context prioritises avoiding false auto-validations of highly irregular documents, macro recall and the severe NOT→VALID rate are more decision-relevant than overall accuracy in this imbalanced tri-class setting.

4.2 CatBoost performance under oversampling and undersampling

The oversampling–undersampling comparison is used only for model development and is conducted within the training partition. Resampling is applied exclusively to the training folds inside the cross-validation pipeline to avoid contamination. The final performance and all decision-support analyses are reported on the beneficiary-disjoint held-out test set (Sect. 4.3 onward).

Table 4 complements Table 3 by isolating the effect of imbalance-handling within CatBoost under the same leakage-safe, beneficiary-grouped cross-validation protocol (StratifiedGroupKFold on the FIT partition), with resampling performed on the training fold only. This comparison highlights the accuracy–sensitivity trade-off induced by alternative resampling strategies.

Table 4 CatBoost cross-validation performance under SMOTENC oversampling versus random undersampling (training FIT partition; StratifiedGroupKFold with beneficiary grouping; resampling restricted to the training fold)

Metric	Oversampling (SMOTENC) (%)	Undersampling (%)
Cross-validation accuracy	68.78	57.77
F1-score (weighted)	73.18	64.84
Macro recall	49.56	55.09
Precision (validated class)	89.37	91.18
Recall (validated class)	73.39	58.43
False *not* → valid (FN, overall)[1]	2.24	1.48
False *partial* → valid (FN, overall)[2]	5.25	3.38

[1]Percentage of all documents in the evaluation folds that are actually NOT VALIDATED and are predicted as VALIDATED (population-level severe false negative rate)

[2]Percentage of all documents in the evaluation folds that are actually PARTIALLY VALIDATED and are predicted as VALIDATED (population-level false negative rate for partial irregularities)

Overall, the SMOTENC oversampling configuration achieves higher accuracy and better identification of truly VALIDATED expense documents (higher recall for the VALIDATED class), resulting in fewer false alarms on compliant cases. However, this more lenient behaviour is associated with a higher population-level rate of governance-critical false negatives, namely irregular cases that are predicted as fully VALIDATED.

The undersampling configuration, by contrast, is more conservative. It sacrifices overall accuracy and recall for VALIDATED (i.e., it flags more compliant documents for review), but it improves sensitivity to the minority outcomes and reduces reduces the population-level severe NOT VALIDATED → VALIDATED miss rate that is costliest in the verification context. This trade-off motivated the choice of the undersampling-based model for decision-support scenarios.

To avoid confusion under class imbalance, Table 4 reports false negative rates as population-level (overall) percentages, computed over all documents in the evaluation folds; the corresponding class-conditional rates are discussed in Sect. 4.3 to provide a policy-relevant representation of residual risk among minority outcomes.

In practical terms, oversampling produced a model that is more lenient, it avoids burdening auditors with too many false alerts, but it may miss a few irregularities. Undersampling produced a model that is stricter, it improves sensitivity to irregular outcomes, but many perfectly valid expense documents would also be flagged for review (false positives). To decide between these, the cost of each error type was considered in the audit context:

- A false negative (failing to flag a problematic expense document, i.e. predicting "valid" when it was not) means an ineligible expense could be reimbursed, resulting in a direct financial loss and undermining the fund's integrity.
- A false positive (flagging an expense document as high-risk when it is actually fine) means an unnecessary audit check is performed. This incurs a cost in staff time and effort but does not directly result in fund losses.

From a risk management perspective, false negatives are far more costly than false positives. It is preferable to err on the side of caution, i.e. review some extra expense documents that turn out to be valid, rather than miss an expense document that should have been rejected. A preliminary cost analysis with the Managing Authority reinforced this view: the potential financial loss from undetected irregular expense documents greatly exceeded the audit cost of examining a few additional documents.

Consequently, the undersampling strategy was selected for the final model. The drop in overall accuracy was deemed an acceptable trade-off for the substantial reduction in undetected errors. In practice, this means the training data used for the final CatBoost model was the one where the VALIDATED class had been down-sampled. It is acknowledged that if circumstances were different (for example, if audit resources were extremely limited and only a very small fraction of expense documents could be reviewed), one might favor a more precise model with fewer false positives. In this case, the priority was clearly to minimize undetected issues, given that 100% of expense documents were traditionally being checked and the aim was to safely reduce this workload. The external (test-set) performance of this model is reported in Sect. 4.3.

Table 5 Precision-oriented confusion matrix (column percentages) on the held-out test set under the calibrated threshold policy

Actual outcome	Predicted VALIDATED (%)	Predicted PARTIALLY VALIDATED (%)	Predicted NOT VALIDATED (%)
VALIDATED	90.59	79.43	84.39
PARTIALLY VALIDATED	7.78	17.75	4.62
NOT VALIDATED	1.63	2.82	11.00

Table 6 Recall-oriented confusion matrix (row percentages) on the held-out test set under the calibrated threshold policy

Actual outcome	Predicted VALIDATED (%)	Predicted PARTIALLY VALIDATED (%)	Predicted NOT VALIDATED (%)
VALIDATED	44.22	37.04	18.74
PARTIALLY VALIDATED	28.99	63.19	7.82
NOT VALIDATED	17.44	28.92	53.64

Table 7 Per-class performance on the held-out test set under the calibrated threshold policy (Wilson 95% confidence intervals for precision and recall)

Outcome class	Precision [95% CI]	Recall [95% CI]	F1	Support (n)
NOT VALIDATED	0.110 [0.098, 0.124]	0.536 [0.490, 0.582]	0.183	453
PARTIALLY VALIDATED	0.178 [0.167, 0.189]	0.632 [0.605, 0.658]	0.277	1304
VALIDATED	0.906 [0.897, 0.914]	0.442 [0.433, 0.452]	0.594	9954

4.3 Final model performance and operational implications

The final CatBoost classifier was estimated on beneficiary-disjoint training data using the tuned hyperparameters and evaluated on a held-out test set containing only beneficiaries not observed during training. Because the split is beneficiary-grouped, test-set performance provides a leakage-safe—and typically more conservative—estimate than the document-level cross-validation figures used during model selection (see Tables 3 and 4). Test-set performance under the calibrated operational policy is summarised in Tables 5, 6 and 7. Table 5 reports the column-normalised (precision-oriented) confusion matrix, Table 6 reports the row-normalised (recall-oriented) confusion matrix, and Table 7 provides per-class precision/recall/F1 with Wilson 95% confidence intervals for precision and recall. Given the audit context, particular attention is placed on the severe error NOT VALIDATED → VALIDATED, which corresponds to an inappropriate auto-validation of a highly irregular expense document.

Under the threshold policy, documents are auto-validated when the predicted probability of the VALIDATED class satisfies, $\hat{P}(VALIDATED) \geq 0.50$; this auto-validation threshold was selected on a beneficiary-disjoint calibration set using post-hoc calibrated probabilities. Under this policy, 41.5% of test documents are auto-validated and 58.5% are routed to manual review. The governance-critical error (NOT VALIDATED → VALIDATED) equals

17.44% conditional on the NOT VALIDATED class (79/453), corresponding to 0.67% of all test documents (79/11711). Conditional on auto-validation (predicted VALIDATED), 90.59% of decisions correspond to truly VALIDATED documents, while 9.41% are irregular outcomes (7.78% PARTIALLY VALIDATED and 1.63% NOT VALIDATED).

Predictions of PARTIALLY VALIDATED and NOT VALIDATED are used as conservative triage labels within the manual-review queue. Their precision is intentionally modest under a risk-averse operating point: within PARTIALLY VALIDATED predictions, 17.75% are genuine partial irregularities and 2.82% are full rejections, whereas 79.43% are ex post fully compliant; among documents predicted as NOT VALIDATED, 11.00% are ultimately rejected and 4.62% are partially validated, while 84.39% are validated (Table 5, column-normalised). The policy therefore accepts a high false-alarm rate in the manual-review queue in exchange for tighter control of false auto-validations.

Aggregate metrics under the threshold policy reflect the deliberate cost asymmetry between false auto-validations and additional manual checks. Macro-average recall equals 53.7% and weighted F1 equals 54.3% (Table 7). From a binary governance perspective (irregular = NOT/PARTIALLY VALIDATED), approximately 74% of irregular documents are routed to manual verification on the test set; the remaining ~26% are auto-validated and constitute the residual exposure controlled by the chosen threshold. Importantly, reporting the severe error both overall (0.67% of all documents) and conditional on the NOT VALIDATED class (17.44%) provides a transparent representation of residual risk under class imbalance.

Operationally, the model is coupled with a threshold policy on $P(\text{VALIDATED})$, which acts as an interpretable control knob linking confidence, workload capacity and residual risk. The auto-validation threshold is selected on a beneficiary-disjoint calibration set via a grid search over candidate thresholds (0.50–0.95), maximising the auto-validation rate subject to a severe-error constraint (NOT VALIDATED → VALIDATED $\leq 2\%$) and using macro-average recall as a secondary tie-breaker. For reference, the unconstrained argmax decision rule would auto-validate 43.3% of test documents (those predicted as VALIDATED), with a severe NOT VALIDATED → VALIDATED rate of 0.83%.

To quantify the marginal contribution of the composite beneficiary-level indicator (beneficiary_risk_score) and mitigate potential circularity concerns, an ablation experiment re-estimates the CatBoost model without this predictor, while keeping the remaining leakage-safe training, calibration, and evaluation protocol unchanged. Table 8 reports test-set results under the post-calibration threshold policy, showing comparable performance and similar minority-class sensitivity.

Table 8 Ablation test on the held-out test set (post-calibration threshold policy): full model versus model without beneficiary_risk_score

Setting (TEST, post-calibration policy)	Macro recall	Weighted F1	Recall NOT	Recall PARTIAL	Recall VALID	Severe NOT→VALID (overall)	Severe NOT→VALID (conditional on NOT)
Full model (with beneficiary risk score)	0.5369	0.5431	0.5364	0.6319	0.4422	0.00675	0.1744
Ablation (without beneficiary risk score)	0.5292	0.5175	0.5386	0.6380	0.4110	0.00581	0.1501

The ablation results indicate that beneficiary_risk_score provides incremental signal but does not solely drive minority-class detection under the leakage-safe protocol. While thresholding provides a direct auto-validation rule, audit planning is often constrained by capacity expressed as a budget on the share of documents that can be fully verified. Table 9 therefore evaluates a complementary ranking-based policy in which expense documents are ordered by an expected-loss proxy defined as the reported amount multiplied by the predicted probability of an adverse outcome, $P(\text{NOT VALIDATED}) + P(\text{PARTIALLY VALIDATED})$. For a given audit budget (i.e., the fraction of documents fully verified), the top-ranked cases are selected for verification and the remainder are auto-validated. Table 12 extends the same back-testing framework to the full workload–risk frontier across audit budgets.

Under this counterfactual policy, the residual (undetected) irregular amount on the held-out test set equals €2.214 million (1.51%) when auditing the top 20% of cases, and decreases to €0.959 million (0.66%) when auditing the top 40%. These figures quantify the marginal risk reduction associated with additional audit capacity and provide a transparent basis for selecting an operating point.

In consultations with the Managing Authority, auditing 40% of cases was assessed as operationally feasible. The released capacity from the remaining 60% can be reallocated to preventive guidance, targeted follow-up on complex projects, or higher-depth checks, illustrating how probabilistic risk scores enable proportional controls. In deployment, audit intensity can be adjusted by revisiting the audit budget and the ranking/thresholding rules in light of evolving capacity constraints and risk tolerance, while monitoring probability calibration and temporal drift.

Moving to model interpretability, a feature importance analysis was performed on the final CatBoost model. Figure 1 highlights how the CatBoost model distributes predictive weight across the available predictors.

The ranking in Fig. 1 confirms that information directly related to the monetary magnitude of an expense document dominates the decision process: *reported_amount_doc* alone explains more than 13% of the total gain, followed by the overall amount reported in the declaration (*reported_amount_declaration*) and the budgeted reference value for the specific cost item (*budgeted_amount_item*). A fourth financial measure, the total project budget used to calculate the aid (*eligible_amount_project*), appears in the mid-range of the ranking, contributing roughly 3% of the gain. Variables that describe the administrative nature of the expense document, most notably *doc_type* and *declaration_type*, also rank highly, indicating that the formal classification of a document provides a strong first signal of risk. The

Table 9 Counterfactual policy back-testing on the held-out test set: undetected irregular amount under illustrative audit budgets using top-k ranking by an expected-loss proxy (reported amount × adverse-outcome probability)

Audit policy	Documents audited (%)	Undetected irregular amount on the test set (€)	Undetected amount as percentage of total reported expenditure (%)
Top-k ranking (expected-loss proxy: reported × risk)	20	2.214 million	1.51
Top-k ranking (expected-loss proxy: reported × risk)	40	0.959 million	0.66

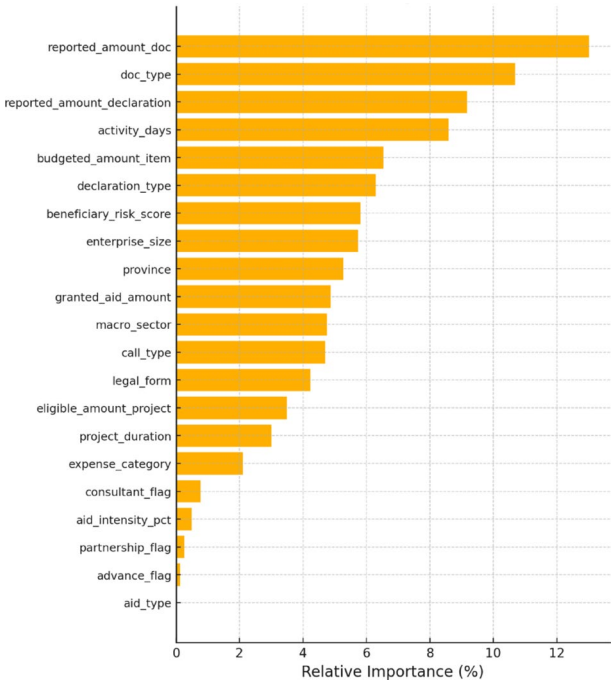

Fig. 1 Relative importance of the predictor variables (described in Table 1) in the final CatBoost model, evaluated on the test set. Bars report each feature's contribution to the total gain, as measured by CatBoost's built-in loss-function-change importance, expressed as a percentage

broad cost grouping captured by *expense_category*, although lower in relative importance (about 2%), still adds useful discriminative power across expenditure classes.

Beyond pure amounts and document descriptors, several contextual factors receive substantial weight. The beneficiary-level composite indicator (*beneficiary_risk_score*), the firm's *enterprise_size*, and the *province* of the beneficiary together account for roughly 16% of total importance, confirming that historical reliability and territorial context remain informative even after controlling for financial details. Project characteristics contribute as well: *activity_days* and *project_duration* capture implementation intensity and timeline, while *granted_aid_amount* reflects the size of the public contribution. Sectoral and procedural attributes, *macro_sector*, *call_type*, and *legal_form*, each provide around 4–5% of gain, suggesting that sector-specific regulations and organisational form affect error likelihood.

At the lower end of the scale, flags such as *consultant_flag*, *aid_intensity_pct*, *partnership_flag*, *advance_flag* and *aid_type* register importance values below 1%. These variables still enter the model but their marginal contribution is limited once the dominant predictors are considered.

It is stressed that variable importance reflects relative influence, not the direction of the relationship; a higher value does not imply that larger numerical values raise or lower the estimated risk. Interpreting sign and shape effects would require additional tools, such as SHAP (SHapley Additive exPlanations) value decomposition or partial-dependence analysis (Molnar 2020), which lie outside the present scope. Nonetheless, the prominence of amount-related and document-type variables aligns with control-authority intuition, providing further reassurance that the model captures meaningful risk patterns in expenditure-verification data.

In terms of error analysis, some of the largest remaining discrepancies were analysed. A few documents that were actually NOT VALIDATED but were predicted as PARTIALLY VALIDATED were studied. These tended to be borderline cases – for instance, a document with a medium-sized issue that the auditor ultimately decided was serious enough to reject entirely. The model, lacking the full context or perhaps specific rule-based nuances, grouped it with more minor irregularities. This suggests that incorporating additional data (for example, text descriptions of findings or specific rule violations) could further help the model distinguish between partial and full rejections. Conversely, there were cases where the model predicted NOT VALIDATED but the outcome was only PARTIALLY VALIDATED. These are essentially over-alarms (the model being stricter than the auditor). They are less concerning from a risk perspective, as they still result in a flagged review, but they do indicate areas where the model could potentially be calibrated to be less sensitive.

Thus, the evidence indicates that the proposed predictive model can stratify expense documents by verification risk and support proportional control strategies. In particular, prioritising verification effort toward higher-risk cases—either via probability thresholding for auto-validation or via ranking-based allocation under a fixed audit budget—can reduce manual workload relative to exhaustive checking while limiting the residual exposure associated with false auto-validations. At the same time, sensitivity to minority outcomes remains moderate and shows temporal variability, motivating conservative governance (threshold setting, monitoring, and periodic refresh) when the model is deployed in changing administrative environments. Although illustrated on EU-funded expenditure data, the modelling and evaluation framework is domain-agnostic and can be transferred to other verification or inspection settings characterised by skewed ordinal outcomes.

4.4 Temporal robustness, calibration, and decision-analytic sensitivity

To probe temporal drift, an expanding-window rolling-origin evaluation was conducted. To prevent within-beneficiary leakage across time (i.e., the same beneficiary appearing in both training and test across different years), the evaluation enforces strict beneficiary separation in each fold: for each test year t, the model was trained on all observations dated prior to t, while the test set included only expense documents in year t belonging to beneficiaries whose first appearance in the dataset occurs in year t ("new beneficiaries only"). Performance is reported under the argmax decision rule, including macro recall, weighted F1, class-specific recalls, and the severe error rate (NOT VALIDATED → VALIDATED), both overall and conditional on NOT VALIDATED (Table 10). Macro recall ranges from 0.461 (2018) to 0.643 (2022), while NOT VALIDATED recall varies markedly (0.128 in 2023 vs. 0.847 in 2020), suggesting that minority-class separability is the main source of temporal instability. In contrast, VALIDATED recall is comparatively more stable across years. The results exhibit non-negligible variability across years, consistent with changes in administrative practice, beneficiary behaviour, and documentation quality. This motivates routine monitoring of performance and periodic model refresh cycles when the operating environment shifts. Because years with a small number of new beneficiaries may display higher sampling variability (e.g., 2023), year-by-year results should be interpreted as indicative of drift rather than as definitive performance guarantees.

A compact visual summary of the rolling-origin evaluation is provided in Appendix A (Fig. 4).

Table 10 Leakage-safe rolling-origin time validation (train on years < t; test on year t for "new beneficiaries only", i.e., beneficiaries whose first appearance year equals t): performance by test year (argmax decision rule)

Test year	*n* (docs)	*n* (beneficiaries)	Macro recall	Weighted F1	Recall NOT	Recall PARTIAL	Recall VALID	Severe overall	Severe cond. on NOT
2018	5220	295	0.461	0.641	0.579	0.145	0.661	0.0285	0.410
2019	3693	536	0.464	0.509	0.729	0.218	0.445	0.0428	0.247
2020	3357	185	0.530	0.621	0.847	0.173	0.570	0.0051	0.124
2021	2294	204	0.612	0.592	0.603	0.764	0.469	0.0083	0.302
2022	2005	138	0.643	0.524	0.765	0.761	0.403	0.0095	0.165
2023	798	38	0.411	0.412	0.128	0.740	0.367	0.0739	0.444

The number of test beneficiaries refers to beneficiaries first observed in the corresponding test year; beneficiaries seen in earlier years are excluded from the test set to avoid leakage across time

Table 11 Probability calibration diagnostics (Pre vs. Post temperature scaling) on the beneficiary-disjoint CALIB subset and on the held-out TEST set

Set	Stage	Multiclass Brier	ECE (max-prob, 10 bins)
CALIB	Pre	0.5148	0.0439
CALIB	Post (temperature scaling)	0.5157	0.0481
TEST	Pre	0.5678	0.0878
TEST	Post (temperature scaling)	0.5699	0.0874

Given that the operational rules in Sect. 4.3 rely on the probability scale, calibration is assessed and corrected within the study. Post-hoc calibration is implemented via temperature scaling fitted on a beneficiary-disjoint calibration subset (Guo et al. 2017). The optimal temperature ($T = 0.954$) is estimated by minimising the multiclass negative log-likelihood on CALIB. Table 11 reports calibration quality before and after temperature scaling on both CALIB and TEST using the multiclass Brier score and the Expected Calibration Error (ECE). On TEST, ECE changes from 0.0878 (pre) to 0.0874 (post), indicating no material deterioration and a slight improvement in aggregate calibration. Threshold selection remains stable (the auto-validation probability threshold remains 0.50), while the implied auto-validation rate changes marginally. Accordingly, all probability-threshold policies and the workload–risk frontier are computed on the calibrated probability scale. Figure 2 complements the tabular diagnostics by reporting one-vs-rest reliability diagrams on TEST before and after temperature scaling.

Calibration is summarised via the multiclass Brier score and the Expected Calibration Error (ECE; max-probability binning, 10 bins). The dashed diagonal indicates perfect calibration; deviations indicate over- or under-confidence across probability bins.

Building on the illustrative operating points in Table 9, the counterfactual back-test is extended by varying the audit budget from 10 to 60% of documents fully verified and computing the resulting workload–risk frontier under a generic top-k policy. Documents are ranked by an expected-loss proxy (reported amount × adverse-outcome probability, $P(\text{NOT VALIDATED}) + P(\text{PARTIALLY VALIDATED})$); the top-ranked fraction is

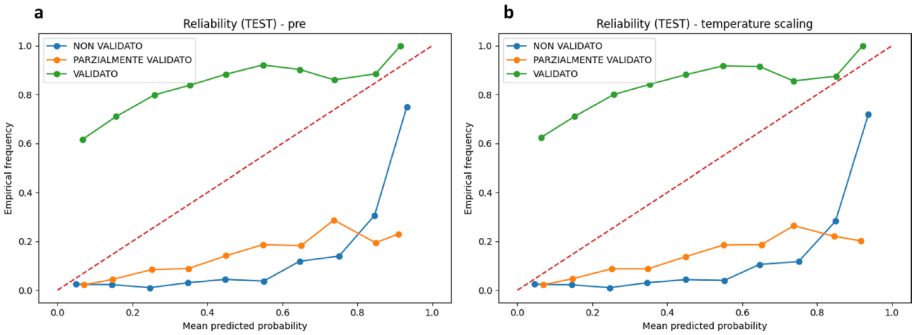

Fig. 2 One-vs-rest reliability diagrams on the held-out test set for the final CatBoost model: **a** uncalibrated probabilities (pre), and **b** probabilities after temperature scaling (post)

Table 12 Workload-risk frontier on the held-out test set: residual (undetected) irregular amount under alternative audit budgets (top-k ranking by expected-loss proxy: reported amount × adverse-outcome probability)

Audit budget (docs) (%)	Undetected irregular amount (€ million)	Undetected as % of total reported
10	2.88	1.97
15	2.49	1.70
20	2.21	1.51
25	1.72	1.18
30	1.35	0.92
35	1.08	0.74
40	0.96	0.65
45	0.78	0.53
50	0.60	0.41
55	0.47	0.32
60	0.31	0.21

audited, while the remainder is auto-validated. For each budget, the residual (undetected) irregular amount is computed as the total irregular expenditure that would remain in the auto-validated set under that policy (using realised outcomes in the held-out test set). Table 12 reports the resulting frontier, providing a transparent mapping between capacity constraints and residual exposure and supporting selection of an operating point. A graphical representation of the workload–risk frontier is provided in Appendix A (Fig. 3).

The frontier is strictly decreasing but exhibits diminishing returns, with the marginal reduction in residual exposure tapering beyond mid-range budgets. For example, increasing the audit budget from 20 to 40% reduces residual exposure from €2.21 million (1.51%) to €0.96 million (0.65%), whereas moving from 40 to 60% yields a smaller absolute reduction (to €0.31 million, 0.21%). This pattern suggests that moderate budgets capture most of the recoverable risk, while higher budgets mainly reduce the tail of residual exposure.

5 Discussion

This study suggests that a supervised-learning engine can translate the regulatory requirement for risk-based verification into a practically usable decision-support routine, subject to validation and governance safeguards. By assigning each incoming expense document a probabilistic risk score, the model supports replacing exhaustive 100% checks with a proportionate audit plan. In the held-out back-test, auditing the upper two-fifths of documents ranked by adverse-outcome probability (or expected-loss proxy) reduces residual undetected irregular expenditure compared to lower audit budgets, while substantially reducing manual workload compared to exhaustive verification. Consistent with the workload–risk frontier, marginal gains diminish beyond mid-range budgets (e.g., from 40 to 60%). When resources are scarcer, the same framework can tighten selection by adjusting the budget or probability thresholds, with a transparent trade-off in residual undetected irregular amount.

Managerial implementation centres on three pillars. First, the predictive module can be embedded in the case-management system to automatically generate risk scores at intake, subject to local IT integration and data availability. Second, governance rules must translate continuous probabilities into actionable decisions through threshold logic or tiered schemes, reviewed periodically to reflect capacity constraints and evolving regulations. Third, a retraining protocol ensures that the model adapts to new data and emerging risk patterns, preserving continuity in the control system. Together, these elements institutionalise risk scoring as part of routine verification rather than a one-off analytical exercise.

The variable-importance profile (Fig. 1) reinforces the model's face validity. Features linked to monetary magnitude, *reported_amount_doc*, *reported_amount_declaration*, *budgeted_amount_item*, account for about one third of total gain, confirming that larger expense documents warrant closer scrutiny. Administrative descriptors (*doc_type*, *declaration_type*) also rank highly, indicating that certain formats or procedural stages carry systematically higher risk. Contextual signals such as *beneficiary_risk_score*, *enterprise_size* and *province* together contribute a further 16%, suggesting that organisational reliability and territorial factors still matter after controlling for financial size. Conversely, procedural flags (*consultant_flag*, *partnership_flag*, *advance_flag*) display marginal influence once the dominant predictors have been considered. The alignment of these findings with auditors' domain knowledge eases adoption and mitigates the "black-box" concern (i.e., the perception that model logic is opaque to auditors) often associated with artificial-intelligence tools in the public sector.

The approach is transferable to other Operational Programmes and Member States, as only local data and a parameter file require adaptation. Transparency can be further strengthened by routinely publishing aggregate feature-importance plots and confusion-matrix summaries, demonstrating that the model supports rather than replaces professional judgement. Embedding the ranking into existing sampling manuals, combined with a documented "four-eyes" principle, addresses fairness and accountability considerations and aligns with current guidance on trustworthy AI.

While the results highlight the practical promise of the proposed framework, some considerations qualify its broader applicability. The evidence comes from a single regional programme, and transferability to other contexts or programming periods requires further validation. In addition, the hierarchical nature of expenditure data suggests that group-aware and time-aware evaluation will remain essential to avoid overstating performance.

The leakage-safe rolling-origin evaluation (new beneficiaries only) shows non-negligible year-to-year variability, suggesting that monitoring and periodic refresh are necessary when administrative processes drift. Any operational deployment would also need suitable governance arrangements, such as documented decision rules, periodic checks for model drift and calibration, and human oversight, to ensure accountability and proportionality. These factors outline the conditions under which the approach can be responsibly scaled in routine verification practice.

6 Conclusions

This study develops and empirically evaluates a supervised-learning engine for tri-class verification outcomes using administrative micro-data from the POR FESR 2014–2020 and a CatBoost-based classifier. The results indicate that selective verification strategies could reduce manual workload while limiting undetected irregular expenditure, provided that the model is externally validated and embedded in a governed decision process.

The research contributes to performance-management theory by tackling an ordered and severely imbalanced classifica. tion problem that defeats conventional statistical models; it also illustrates how model choice should be guided by the asymmetric cost of errors, as undetected irregularities are far more harmful than false alarms. On the managerial side, the workflow translates abstract guidance on "proportionate controls" into explicit decision rules. Feature-importance results confirm that the algorithm relies on variables already familiar to auditors, claimed amount, document type, beneficiary risk and provincial context, thereby supporting rather than replacing professional judgement. When integrated into existing declaration-processing workflows, the model can provide ex ante risk scores to support triage decisions, subject to local IT integration and governance safeguards.

Methodologically, the study offers a replicable blueprint: rigorous preprocessing, CatBoost training, transparent importance ranking and leakage-safe evaluation designs (beneficiary-grouped splits and time-based validation), complemented by a scheduled retraining protocol. Only the local data set and a parameter file need adaptation, making the approach transferable to other programmes and Member States.

Several limitations should nonetheless be acknowledged. The model was calibrated on a single regional ERDF programme, and external validity should be established through multi-site replication. Temporal robustness, assessed via a leakage-safe rolling-origin evaluation on "new beneficiaries only", exhibits non-negligible year-to-year variability, suggesting that monitoring and periodic refresh are necessary under drift. The dataset comprises only structured fields; incorporating unstructured evidence (e.g., narrative justifications) may capture additional risk cues. The interface currently provides global probabilities; adding SHAP-based local explanations could extend transparency and auditability. Finally, the current architecture treats partial and full validation failures symmetrically, whereas some authorities may wish to prioritise the most severe irregularities.

Future research should therefore explore cross-regional experiments to assess generalisability; cost-sensitive learning to optimise expected monetary loss; active-learning strategies to refine predictions on borderline cases; and the integration of predictive scores with process-mining techniques to support earlier, preventive interventions. In addition, further work could compare alternative post-hoc calibration methods and assess their effect on

threshold-based policies in operational settings. These directions can further enhance the tool's effectiveness and contribute to the broader agenda of data-driven performance management in public finance.

Overall, the findings show that data-driven triage can reconcile regulatory expectations with resource constraints, enabling audit authorities to improve both efficiency and effectiveness while fostering a culture of continuous, evidence-based improvement. Beyond its domain relevance, the study contributes a generalisable quantitative blueprint for imbalanced ordered prediction on administrative data, aligning with ongoing efforts to integrate advanced statistical and ML methods into decision-support processes.

Appendix: supplementary decision-support diagnostics

To support transparency, this appendix provides complementary diagnostics that summarise the trade-off between audit capacity and residual risk, and the stability of performance over time. Figure 3 visualises the workload–risk frontier on the held-out test set under a top-k ranking policy based on the expected-loss proxy (reported amount × adverse-outcome probability). Figure 4 summarises the rolling-origin evaluation (train on years $< t$, test on year t), highlighting year-to-year variation in key performance indicators.

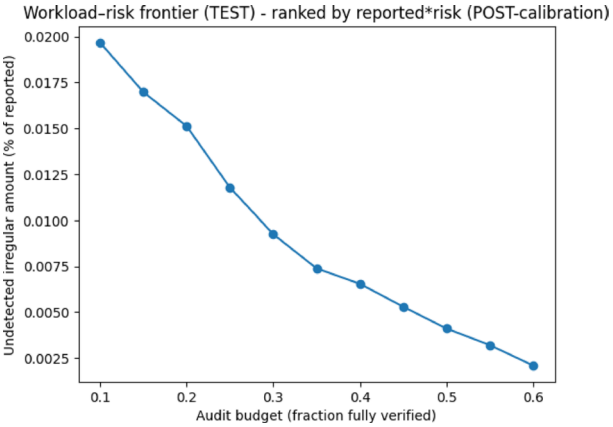

Fig. 3 Workload–risk frontier on the held-out test set (top-k ranking by expected-loss proxy: reported amount × adverse-outcome probability)

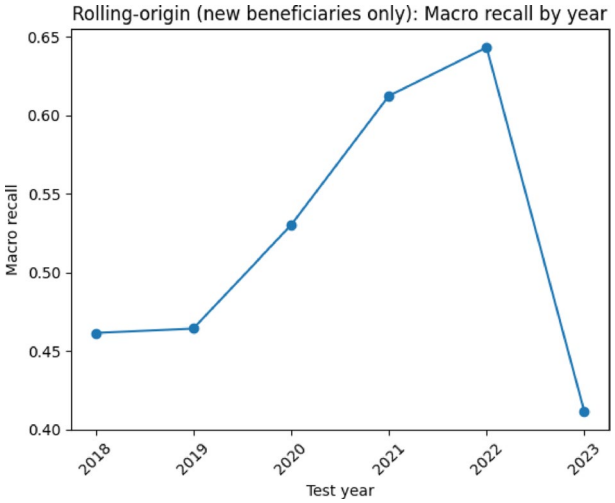

Fig. 4 Rolling-origin time validation summary (train on years $< t$, test on year t)

Acknowledgements The authors gratefully acknowledge Finpiemonte S.p.A. for the valuable collaboration and data provision that made this research possible, in particular Dr. Flavia Stella, Dr. Filippo Marzucchi, and Finpiemonte's Director General Dr. Mario Nicola Francesco Alparone. Their support and domain expertise provided throughout the project were instrumental in aligning the methodological development with real-world operational needs.

Author contributions All authors contributed to the conception and design of the study. E.V. conducted the data analysis and developed the Python code. The manuscript was initially drafted by E.V. and subsequently revised and refined by all authors.

Funding Open access funding provided by Politecnico di Torino within the CRUI-CARE Agreement. This research was carried out with the support of Finpiemonte S.p.A., whose collaboration and contribution are gratefully acknowledged.

Data availability The dataset analyzed in this study was provided by Finpiemonte S.p.A. and is not publicly available due to confidentiality agreements. Derived and anonymized data may be available from the corresponding author upon reasonable request and with permission from Finpiemonte S.p.A.

Declarations

Conflict of interest The authors declare no conflict of interest.

Generative AI and AI-assisted technologies No Generative AI or AI-assisted technologies were used in the conduct of this study or in the preparation of this manuscript. All analyses, results, and interpretations were produced by the authors.

References

Aldana, A., Falcón-Cortés, A., Larralde, H.: A machine learning model to identify corruption in Mexico's public procurement contracts. arXiv preprint (2022). arXiv:2211.01478

Bai, J., Qiu, T.: Automatic Procurement Fraud Detection with Machine Learning. arXiv preprint arXiv:2304.10105. (2023)

Bonnier, T., Bosch, B., Moniz, N., Branco, P., Torgo, L., Japkowicz, N., Wo, M., Wang, S.: Assessing the robustness of ordinal classifiers against imbalanced and shifting distributions. Proc. Mach. Learn. Res. **183**, 112–126 (2022)

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. **16**, 321–357 (2011). https://doi.org/10.1613/jair.953

Ensslin, S.R., Welter, L.M., Pedersini, D.R.: Performance evaluation: a comparative study between public and private sectors. Int. J. Productivity Perform. Manage. **71**, 1761–1785 (2022). https://doi.org/10.1108/IJPPM-04-2020-0146/FULL/PDF

European Commission: Arachne Charter: For the Introduction and Application of the Arachne Risk Scoring Tool in the Management Verifications: (2024)

European Commission: Risk Based Management Verifications - Article 74(2) CPR 2021–2027: (2023)

Galetto, M., Verna, E., Genta, G.: Accurate estimation of prediction models for operator-induced defects in assembly manufacturing processes. Qual. Eng. **32**, 595–613 (2020)

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., Bouchachia, A.: A survey on concept drift adaptation. ACM Comput. Surv. **46**, 1–37 (2014). https://doi.org/10.1145/2523813

Garengo, P., Sardi, A.: Performance measurement and management in the public sector: state of the art and research opportunities. Int. J. Productivity Perform. Manage. **70**, 1629–1654 (2021). https://doi.org/10.1108/IJPPM-03-2020-0102/FULL/PDF

Guignard, F., Ginsbourger, D., Levy Häner, L., Herrera, J.M.: Some combinatorics of data leakage induced by clusters. Stoch. Environ. Res. Risk Assess. 2024. **38**, 7 (2024). https://doi.org/10.1007/S00477-024-02715-1

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: 34th International Conference on Machine Learning, ICML 2017. 3, 2130–2143 (2017)

Halachmi, A., Greiling: D Accountability and organizational performance in the public sector: a symposium introduction. Int. J. Productivity Perform. Manage. **60** 244–251 https://doi.org/10.1108/IJPPM.2011.07960AAA.001/FULL/XML (2011)

Henriquez-Machado, R., Muñoz-Villamizar, A., Santos, J.: Advancing sustainable operational excellence: a machine learning approach for emerging economies. Int. J. Productivity Perform. Manage. ahead-of-print. (2025). https://doi.org/10.1108/IJPPM-05-2024-0332/FULL/PDF

Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. Artif. Intell. Rev. **22**, 85–126 (2004). https://doi.org/10.1023/B:AIRE.0000045502.10941.A9

Kapoor, S., Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science. Patterns. **4**, 100804 (2023). https://doi.org/10.1016/J.PATTER.2023.100804

Kaufman, S., Rosset, S., Perlich, C., Stitelman, O.: Leakage in data mining. ACM Trans. Knowl. Discovery Data (TKDD). **6** (2012). https://doi.org/10.1145/2382577.2382579

Lázaro, M., Figueiras-Vidal, A.R.: Neural network for ordinal classification of imbalanced data by minimizing a Bayesian cost. Pattern Recognit. **137**, 109303 (2023). https://doi.org/10.1016/J.PATCOG.2023.109303

Lemnaru, C., Potolea, R.: Imbalanced classification problems: systematic study, issues and best practices. Lecture Notes in Business Information Processing. 102 LNBIP, 35–50 (2012). https://doi.org/10.1007/978-3-642-29958-2_3

Liang, Z.: Efficient Representations for High-Cardinality Categorical Variables in Machine Learning. arXiv preprint arXiv:2501.05646. (2025)

Molnar, C.: Interpretable machine learning. Lulu. com (2020)

More, A.: Survey of resampling techniques for improving classification performance in unbalanced datasets. arXiv preprint arXiv:1608.06048. (2016)

Murphy, A.H., Winkler, R.L.: Diagnostic verification of probability forecasts. Int. J. Forecast. **7**, 435–455 (1992). https://doi.org/10.1016/0169-2070(92)90028-8

Nawaz, M., Khan, S., Ghulam, ·, Bhatti, A., Muhammad, ·, Khan, J., Nawaz, M., ·, Khan, S.: Evaluating the critical factors of building information modeling implementation using ordinal priority approach and grey relational analysis. Quality and Quantity 1–18 (2025). (2025). https://doi.org/10.1007/S11135-025-02445-8

Nehme, R., AlKhoury, C., Mutawa, A.: Evaluating the performance of auditors: a driver or a stabilizer of auditors' behaviour. Int. J. Productivity Perform. Manage. **69**, 1999–2019 (2020). https://doi.org/10.1108/IJPPM-08-2018-0306/FULL/PDF

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2012)
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. **2018**, 6638–6648 (2017)
Radnor, Z., McGuire, M.: Performance management in the public sector: fact or fiction? Int. J. Productivity Perform. Manage. **53**, 245–260 (2004). https://doi.org/10.1108/17410400410523783/FULL/PDF
Rainio, O., Teuho, J., Klén, R.: Evaluation metrics and statistical tests for machine learning. Sci. Rep. **14**, 1–14 (2024). https://doi.org/10.1038/S41598-024-56706-X
Siciliani, L., Taccardi, V., Basile, P., Di Ciano, M., Lops, P.: AI-based decision support system for public procurement. Inf. Syst. **119**, 102284 (2023). https://doi.org/10.1016/J.IS.2023.102284
van Buuren, S.: Flexible Imputation of Missing Data, Second Edition. Flexible Imputation of Missing Data, Second Edition. (2018). https://doi.org/10.1201/9780429492259
Verna, E., Genta, G., Galetto, M., Franceschini, F.: Inspection planning by defect prediction models and inspection strategy maps. Prod. Eng. Res. Devel. **15**, 897–915 (2021)
Verna, E., Genta, G., Galetto, M., Franceschini, F.: Zero defect manufacturing: a self-adaptive defect prediction model based on assembly complexity. Int. J. Comput. Integr. Manuf. **36**, 155–168 (2023)
Wongkamthong, C., Akande, O.: A comparative study of imputation methods for multivariate ordinal data. J. Surv. Stat. Methodol. **11**, 189–212 (2020). https://doi.org/10.1093/jssam/smab028
Zheng, A., Casari, A.: Feature engineering for machine learning: principles and techniques for data scientists. O'Reilly Media, Inc. (2018)

Authors and Affiliations

Elisa Verna[1] · Gianfranco Genta[1] · Maurizio Galetto[1]

✉ Elisa Verna
elisa.verna@polito.it

Gianfranco Genta
gianfranco.genta@polito.it

Maurizio Galetto
maurizio.galetto@polito.it

[1] Department of Management and Production Engineering, Politecnico di Torino, Torino, Italy