Construction of a CatBoost Classification Prediction Model for Municipal-Level Teacher Identity Based on Professional Experience

Jingyi Shan, Linan Lin*

College of Teacher Education, Shenyang Normal University, Shenyang, China
**Corresponding Author. Email: 18602453365@126.com*

Abstract. To examine the relationship between teacher identity characteristics and professional experience, the present study uses municipally recognized outstanding teachers in the Beijing-Tianjin-Hebei region as its research sample and develops a teacher identity classification prediction model using the CatBoost machine learning algorithm. Under the guiding premise of "promoting the great spirit of educators," 14 professional-experience feature indicators are extracted from four dimensions, including educational and instructional competence, collaborative and innovative capacity, and research-and-practice capability, among others. By drawing on CatBoost's well-established advantage in processing categorical features, the analysis estimates the relative importance of different identity characteristics among outstanding teachers and incorporates those characteristics into variable training and testing procedures. On the methodological front, CatBoost improves on the conventional Gradient Boosting Decision Tree (GBDT): through evaluation on test data, classification performance is sharpened, and model assessment, in turn, is rendered more precise. The findings show that machine learning is broadly applicable to the evaluation of teacher professional development and can accurately disclose the structural composition of teacher professional identity. That, in practical terms, supplies data-based support and scientifically grounded evidence for advancing teacher professional development, designing targeted training strategies, and furthering the construction of a strong education system.

Keywords: Municipal-Level Teachers, CatBoost Classification Prediction, Identity Characteristics

1. Introduction

The Education Powerhouse Development Plan (2024–2035) places renewed emphasis on the strategic objective of "vigorously promoting the spirit of educators and cultivating a high-quality teaching workforce." Set against the still-shifting demands of educational modernization, conventional teacher education models have become progressively less capable of meeting the requirements associated with building an educational powerhouse. Two deficiencies stand out, and they are not trivial ones: first, the uneven regional distribution of high-quality teacher resources; second, the absence of a scientific evaluation system for teacher professional development. Under these conditions, educational administrative departments and researchers in teacher education are required to draw on contemporary information technologies—notably big data and artificial intelligence—to identify and screen the core characteristics of teacher professional development, thereby pushing teacher education paradigms toward transformation and structural upgrading. From that policy and practical juncture, the construction of a dynamic evaluation model for teacher professional development through machine learning algorithms has emerged as a central concern in deepening reform of the teacher education system.

For a long period, traditional teacher training has depended primarily on theoretical coursework and the experience-based transmission of knowledge from mentors to trainees. In practice, however, the limitations of that arrangement have become difficult to ignore. To begin with, theory and practice do not connect particularly well. Pre-service teacher preparation still centers mainly on educational theory, with only short-term teaching internships added on, and prior studies suggest that pre-service teachers frequently display limited awareness of professional development together with insufficient job competence, which in turn prolongs their adjustment period after entering professional posts. A further weakness lies in the homogenization of course content and, relatedly, the neglect of individual differences among teachers. Existing teacher education curricula do not adequately respond to the personalized developmental needs of teachers across disciplines; especially under-addressed are the distinctive needs of special education teachers (a gap long noted in practice, though not always built into training design). No less constraining, the evaluative framework remains reductively narrow: judgments of teacher excellence depend too heavily on fixed markers such as academic credentials and the sheer number of honors or awards conferred. What remains absent—or, more precisely, insufficiently captured—are credible indicators of more fluid professional capacities, including digital literacy, research productivity, and instructional innovation. From this has emerged a markedly instrumental orientation, professional development being pursued chiefly to secure higher ranking positions rather than to foster substantive professional growth. Beyond that, support for teachers' lifelong development remains inadequate. Professional development does not end with initial teacher preparation; instead, it continues across the entirety of a teacher's career trajectory. For that reason, schools and educational authorities must renew their training philosophies, put in place an integrated model of teacher development, and furnish differentiated platforms and resources more appropriately aligned with sustaining teachers' professional growth at distinct career stages.

Over the past several years, the artificial intelligence industry has expanded at pace, and educational data mining has been reshaped in tandem. Through the construction of analytic models, the field has started to move across—and, in part, circumvent—longstanding assessment constraints, thereby improving evaluative efficiency and, at the same time, prompting more inventive uses of machine learning algorithms in teacher education. Once dynamic feature extraction, nonlinear relationship modeling, and the coupled logics of prediction and attribution are brought into a single analytical frame, intelligent, standardized assessment of teacher development becomes feasible. On the empirical side, the record reported thus far suggests that evaluation models built on machine learning outperform traditional approaches by a clear margin, particularly in predictive speed and accuracy. Consider, for instance, the unsupervised Random Forest (RF) approach: using the Gini index, permutation importance, and the Boruta algorithm, it has been applied effectively both to ranking the importance of cloud computing-related variables [1] and to identifying and analyzing intelligent strategies associated with family education models [2]. On that groundwork, the present study introduces the CatBoost machine learning classification algorithm and takes the professional experience of municipal-level teachers as its analytic basis. By extracting discriminative temporal features from categorical data, an interpretable model consistent with disciplinary logic is constructed, making possible the continuous monitoring and attribution analysis of teacher growth potential. In more compact terms, the study extends the research paradigm for teacher professional development.

2. Literature review

Within the scholarship on teacher professionalism, a robust disciplinary grounding together with strong instructional capacity constitutes a basic—indeed study-defining—dimension of excellent teachers' professional competence. On one side of this dual requirement, outstanding teachers are expected to command deep subject knowledge. Beyond mastering foundational educational theory, they must also develop a curriculum-attuned, conceptually rich grasp of the goals and core ideas embedded in their disciplines, and then translate that subject-specific knowledge system into actual classroom practice in an effective manner [3]. Set alongside this, excellent teachers display a marked capacity to recalibrate teaching methods and instructional strategies in light of students' differing interests and learning needs; such adaptability, far from being incidental, heightens student engagement and encourages active learning. In the New Basic Education Reform experiment, Professor Lan Ye explicitly advanced the proposition of "returning the classroom to students," arguing that the classroom should function as a site of students' autonomous learning and active inquiry. That pedagogical orientation has been put into practice by outstanding teachers who place students at the center of learning, continue to revise and innovate both classroom teaching and classroom-management approaches, and realize an organic coupling of teacher guidance with student autonomy.

At the level of professional formation, conviction operates as the animating force behind teachers' sustained advancement and remains fundamental to their development. A noble educational belief signals not only a teacher's commitment to education, but also an accompanying sense of social responsibility. From this follows a practical impetus: teachers are motivated to explore new educational approaches so as to cultivate students with the essential character qualities and key competencies required for adaptation to future social development and for lifelong personal growth [4]. Over recent years, against the backdrop of increasing emphasis on the spirit of educators, the importance of teacher ethics within the teacher-education system has become more visible. Teacher ethics, understood as the behavioral norms and moral standards teachers are obliged to uphold while carrying out professional duties, encompass more than commitment to the education sector alone; included as well are dedication and selflessness. The saying "A teacher must be learned to teach and upright to be a model" captures, with particular economy, teachers' dual identity as disseminators of knowledge and guides in students' personal development. By exemplifying noble moral character and, at the same time, cultivating students through profound professional knowledge, teachers embody the unity of the academic instructor and the moral educator.

For teachers moving toward research-oriented professionalism, innovative awareness and digital competence operate not as optional enhancements but as prerequisite conditions [5]. Against the policy framework established by China Education Modernization 2035, markedly higher requirements have been imposed on teachers' digital literacy and on their ability to deploy educational technology in effective classroom use. With the arrival of the artificial-intelligence era, the education sector has, in actual practice, begun to witness the emergence of a new generation of educators—innovative in orientation and notably future-aware. Rather than remaining bound to conventional instructional models, these teachers take up AI technologies proactively and integrate them into classroom practice, thereby producing pedagogical approaches that are not simply new in form but demonstrably effective, with corresponding improvements in both instructional quality and instructional efficiency.

Beyond subject-matter command alone, what sets exemplary teachers apart is the capacity to adjust to technological change and to use educational technologies to construct learning platforms calibrated to individual students. Within this increasingly digitalized—and progressively intelligent—educational environment, differentiated instruction is correspondingly better sustained. Through project-based learning, together with pedagogical arrangements of a similar kind, students are encouraged to apply AI in concrete practice; by doing so, they cultivate innovative thinking and further refine problem-solving capacity, a process that, at the same time, supports educational modernization and the building of a strong education system.

No less central to teachers' professional development are continuous learning and ongoing self-improvement. In a period marked by rapid advancement—knowledge being updated quickly, new technological domains appearing with notable frequency—teachers must sustain an active orientation toward learning. Continuous revision of their knowledge structure is required; so too are the strengthening of teaching competence and the improvement of practical skills, all to remain aligned with broader social change. Teachers committed to lifelong learning can promote their own development, but the matter does not end there: such teachers also engage in ongoing reflection and self-evaluation. Through critical scrutiny of teaching practice, problems can be identified and addressed in a timely way, leading ultimately to improvements in the quality and effectiveness of education.

3. Methodology design

3.1. Data sources

As the underlying data source, this study draws on the professional experience profiles of municipal-level teachers and, on that basis, employs the CatBoost machine learning algorithm to construct a classification prediction model. In line with the operating principles of CatBoost, variables are selected from the professional experience indicator system for municipal-level teachers. More concretely, nine indicators are specified as feature variables: whether practical experience is present, whether the teacher serves as a training expert, participation in research projects, research output, recognition of excellent teaching cases by the JiaoShiWang platform, teacher ranking, collaboration ability, educational background, and administrative experience. Set against these inputs, municipal-level outstanding teacher status is defined as the output variable. The CatBoost model thus constructed estimates the importance of teacher identity characteristics and uses those estimates in both data training and testing. Unlike more conventional models, CatBoost appraises predictive performance through classification-based metrics—accuracy, recall, precision, and F1 score—thereby providing a scientifically grounded assessment of classification performance.

Compared to the traditional Gradient Boosting Decision Tree (GBDT), the CatBoost algorithm is designed to handle categorical features more efficiently. In the processing of categorical variables, cumbersome preprocessing is removed through target encoding, with categorical features being handled directly by means of smoothed mean encoding. Training is therefore accelerated. Over-fitting, at least in part, is also reduced. Turning to sequential tree construction, CatBoost adopts an ordered boosting mechanism, an architectural adjustment under which each tree, during its construction, is permitted to access only the predictions generated by earlier trees.

Owing to its adjustable reliance on multiple loss functions, this approach mitigates the information-leakage problem that recurrently appears in conventional GBDT models (better understood as a methodological limitation than as a purely procedural one). Within CatBoost, a loss function widely used in the literature is specified by two constituent elements: the training error and a regularization term [6]. The loss function is expressed as follows:

$$\mathcal{L}(F) = \sum_{i=1}^{n} L\big(y_i, F(x_i)\big) + \sum_{k=1}^{k} \Omega(f_k)$$

where

$\Omega(f_k)$denotes the regularization term linked to the k-th tree, typically parameterized through the number of leaf nodes and the summed squared leaf weights. Under this penalty specification, unchecked growth in model complexity is curtailed (that is, it is not permitted to expand without restraint), which in turn lowers the risk of overfitting. By contrast, the training error captures the discrepancy between the predicted outputs and the observed values.

To accelerate prediction and preserve robustness under noisy observations or outlier-contaminated cases, CatBoost employs a symmetric tree architecture, a design in which every node within a given tree is partitioned by the same feature-threshold criterion. With that arrangement in place, computational efficiency is maintained and model stability is preserved; in educational data analysis in particular, where measurement irregularities are scarcely rare, that pairing makes the method especially well suited to the task.

4. Empirical process and result analysis

Table 1. Model parameters

Parameter Name	Parameter Value
Training Time	0.085s
Data Split Ratio	0.9
Data Shuffling	Yes
Number of Iterations	100
Learning Rate	0.1
L2 Regularization Term	1
Maximum Tree Depth	10
Over-fitting Detection Threshold	0
Additional Iterations After Optimization	20

As detailed in Table 1, the CatBoost model was configured with the parameter values listed there, and total training finished in 0.085 seconds—a runtime that is, for this specific setup, notably brief. For evaluation, the data-set was divided at a 9:1 training-to-validation ratio, this split having been selected to retain an ample share of data for model fitting while still reserving held-out samples for performance checking.

During preprocessing and the model-fitting stage that followed, shuffling was kept enabled, chiefly to prevent the ordering of samples from exerting undue influence on the learned output; in practical terms, this reduces the chance that sequence-dependent artifacts are carried into the fitted model and, by extension, supports generalizability. Across the iterative optimization routine, the number of iterations was fixed at 100 so that the data could be fit with greater adequacy. The learning rate, held constant at 0.1, functioned meanwhile as a moderating constraint on parameter updates, limiting an optimization path that might otherwise advance too aggressively.

To curb model complexity and, in turn, suppress overfitting, the L2 regularization coefficient was held constant at 1, while maximum tree depth was capped at 10; similar in purpose but not in operation, these two restrictions targeted separate axes of model capacity—penalization in the former instance, tree expansion in the latter. Following completion of the primary optimization phase, an additional 20 iterations were carried out to sharpen performance further. Under this parameter setting, CatBoost's computational efficiency and fitting capacity become plainly visible, strong predictive performance having been attained within a comparatively brief training interval.

Table 2. Feature importance

Feature Name	Feature Importance (%)
Practical Experience	0.00%
Training Expert	3.80%
Research Project Involvement	3.20%
Research Output	42.80%
Recognized Teaching Case (JiaoShiWang)	3.50%
Teacher Ranking	27.70%
Collaboration Ability	0.00%
Educational Background	8.20%
Administrative Experience	10.70%

As reported in Table 2, "Research Output" dominates the model's predictions, accounting for 42.8% of total feature importance and functioning—by a considerable margin—as the primary classification variable. Well behind it, "Teacher Ranking" contributes 27.7%; read jointly, these values indicate that teacher title classification substantively shapes the predicted outcomes. In a more distant third position, "Administrative Experience" accounts for 10.70%, suggesting that this attribute likewise enters the model's decision process in a nontrivial way.

Set against the more heavily weighted predictors, "Educational Background" accounts for 8.2% of the explained contribution, suggesting that academic qualifications remain within the predictive architecture, though with substantially less weight than the variables ranking above it. Closer to the lower end of the contribution spectrum, "Training Expert" (3.8%), "Recognized Teaching Case" (3.5%), and "Research Project Involvement" (3.2%) enter the classification procedure only to a limited extent. Most strikingly, "Practical Experience" and "Collaboration Ability" each register 0.00%, a pattern consistent either with an absence of meaningful incremental predictive value or, alternatively, with the possibility that whatever signal these variables contain has already been subsumed under the model's more influential features.

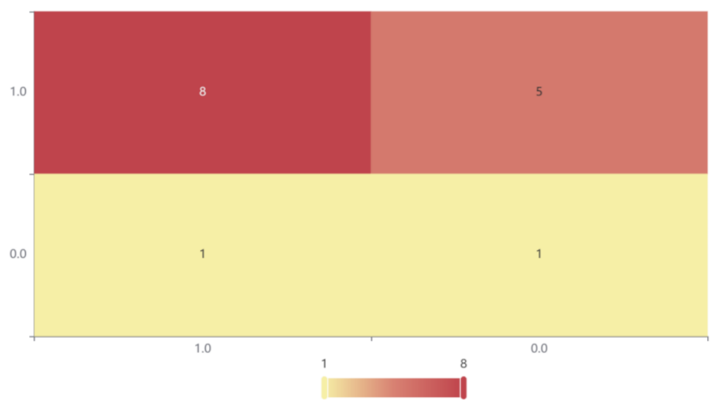

Figure 1. Confusion matrix heatmap

In Figure 1, the confusion matrix presents a distinctly differentiated pattern across its individual cells. Reading the value shown in each position allows the allocation of observations to be traced for every predicted-actual category pairing.

In the top-left cell, the True Positive (TP) count indicates that the model correctly identified one sample as positive. Set against this, the bottom-left cell records the False Positive (FP) count, showing that one negative sample was misclassified as positive. In the top-right cell, by contrast, no analogous error appears: the False Negative (FN) value is 0, meaning that no positive sample was classified as negative. In the bottom-right region of the matrix, the cell with a value of 8, considered alongside the adjacent cell valued at 5, corresponds to the True Negatives (TN), thereby indicating that the model correctly identified most negative samples.

Table 3. Model evaluation results

	Accuracy	Recall	Precision	F1 Score
Training Set	1	1	1	1
Test Set	0.6	0.6	0.793	0.664

Set against the training-phase results, Table 3 throws into especially sharp relief a pronounced divergence between in-sample fit and out-of-sample performance.

Within the training data-set, the CatBoost classification model records perfect metrics—accuracy, recall, precision, and F1 score all equal to 1.000—indicating, under training conditions, classification with essentially no observable error. Once the analysis moves to the test set, however, performance deteriorates in a readily apparent way, which in turn suggests constrained generalizability. On the test data, more specifically, accuracy and recall are each 0.600; put directly, 60% of positive-class observations are classified correctly, and 60% of all truly positive cases are recovered. Against this backdrop, the precision estimate of 0.793 shows that 79.3% of instances labeled as belonging to the positive class are in fact positive—evidence that positive-case predictions remain fairly reliable, though not entirely free from residual misclassification.

5. Conclusion and educational management implications

Drawing on the professional experience of municipal-level teachers, the present study employs the CatBoost machine learning algorithm to construct a classification-based predictive model for identifying the defining characteristics of outstanding teachers. Against the empirical record examined here, CatBoost shows wide-ranging applicability to the evaluation of teacher professional development.

At a more fine-grained level, the analysis isolates several constituent dimensions of the professional competence associated with outstanding teachers. Standing out most clearly are two competencies that sit at the very center of professional-competence formation: a firm grounding in disciplinary knowledge and strong instructional capacity. Beneath these relatively visible dimensions, there operates a deeply internalized sense of professional dedication and ethical responsibility, one that serves as the durable foundation for sustained professional development. Alongside these elements, innovative consciousness—coupled with digital competence—has become a necessary condition for teachers' movement toward research-oriented professionalism. No less apparent is the point that continuous learning, together with sustained self-renewal, remains an indispensable driver of long-term professional growth.

Against the backdrop of these findings, the following implications and recommendations for teacher professional development are set out.

(1) Enhancing Research Awareness:

At the level of teacher development practice, markedly greater emphasis should be placed on cultivating a research-oriented disposition and, alongside that, reaffirming the guiding principle of "advancing education through research." Within this understanding, research is not to be treated as a peripheral task or a detachable add-on; rather, it operates as a practice-embedded mechanism for improving teaching quality and supporting professional growth. Through sustained participation in research projects, academic writing, and scholarly seminars, teachers can deepen their command of core educational concepts—gradually, that is, and in concrete rather than merely abstract ways.

(2) Innovating Teaching Approaches:

Across ordinary classroom practice, teachers should be encouraged to move beyond inherited instructional templates and to test, in a deliberate way, newer pedagogical approaches alongside alternative classroom-management arrangements. Experiential learning, project-based learning, interdisciplinary thematic instruction, and flipped-classroom formats, among other options, can heighten student engagement and learning interest; that shift, in turn, can deepen students' understanding and use of subject knowledge. In contexts where conventional routines constrain responsiveness to learners' needs (a recurrent problem in day-to-day instruction), such innovations may be especially consequential.

(3) Developing a Comprehensive and Multi-Level Teacher Training System:

From the vantage point of educational administration, what is required is the establishment of a layered, system-wide teacher training mechanism capable of sustaining lifelong professional development. With respect to training content, the CatBoost machine-learning algorithm can be deployed to analyze teachers' professional experience and, from that analytical base, distinguish training needs across subject specializations and different phases of career development. On that basis, training curricula may then be calibrated to teachers' actual needs. As programme content is aligned more closely with those needs (rather than being organized around undifferentiated, one-size-fits-all provision), both the quality and the practical efficacy of teacher training can be reinforced; in turn, that closer fit contributes to high-quality teacher development and, at the system-wide level, supports the construction of a strong education system.

References

[1] *Hui, H., Silvana, T. (2024) Analysis of cloud computing-based education platforms using unsupervised random forest. Education and Information Technologies,12,15905-15932.*

[2] *Xia, J., Zhang, S. (2023) Enhancing family education pattern recognition with a random forest algorithm. Journal of Intelligent & Fuzzy Systems,6,9803-9813.*

[3] *Imani, R., Raven, S., Dalia, S., Jamala, W., Hannah, J., René, R., Tiffany, Bergin. (2025) Integration of individuals with lived experience to improve recruitment within criminal justice research: 'experience as the best teacher'. International Journal of Social Research Methodology,1,1-14.*

[4] *Ignacio Jr. A. G., (2024) EXPLORING THE PERSPECTIVES OF PRESERVICE FILIPINO MATHEMATICS TEACHERS: BASIS FOR A PROPOSED EDUCATIONAL BELIEF MODEL. Problems of Education in the 21st Century,4,487-506.*

[5] *Sarkar, C., Mohanty, V., Balappanavar, A., Chahar, P., Rijhwani, K.(2022).Development and Validation of a Comic Tool: An Innovative Approach to Raise Awareness about Tobacco Control among School Teachers. Indian Journal of Community Medicine,4,536-542.*

[6] *Md, E.C. A.K.M., Saiful I.,Rashed,U.Z., Sharfaraj, K.(2025). A machine learning-based approach for flash flood susceptibility mapping considering rainfall extremes in the northeast region of Bangladesh. Advances in Space Research,2,1990-2017.*