

Comparative Evaluation of Ensemble Learning Methods for Capital Expenditure Deviation Prediction in FERC Rate-Base Transmission Projects

Yiyang Peng

Master of Public Administration, University of Southern California, Los Angeles, CA, USA

Keywords

capital expenditure prediction, ensemble learning, FERC transmission regulation, SHAP interpretability

Abstract

Accurate capital expenditure (CapEx) prediction for large-scale transmission projects regulated under the Federal Energy Regulatory Commission (FERC) rate-base mechanism remains a persistent challenge, as cost overruns directly inflate the rates borne by electricity consumers. This study presents a comparative evaluation of ensemble learning methods—XGBoost, LightGBM, CatBoost, and Random Forest—for predicting CapEx deviations in FERC-jurisdictional transmission projects. Drawing on publicly available FERC Form 1 financial data (1994–2019) and supplementary macroeconomic indicators from the Bureau of Labor Statistics (BLS) and the Producer Price Index (PPI) for electrical equipment, we construct a tabular dataset of 1,247 utility-year records encompassing annual transmission capital additions, operations and maintenance expenditures, regional labor indices, and commodity price signals. Each record is labeled with a deviation indicator reflecting whether actual annual capital additions exceeded the prior three-year rolling average by more than fifteen percent. Experimental results across five-fold cross-validation indicate that LightGBM achieves the strongest predictive performance (AUC = 0.841, F1 = 0.762), with XGBoost following closely (AUC = 0.833, F1 = 0.748). SHAP-based feature importance analysis reveals that the construction cost index, year-over-year transmission plant growth rate, and regional wage differential are the three most influential predictors. These findings offer empirical evidence supporting the viability of gradient boosting approaches for early-stage cost deviation screening in regulated transmission investments.

1. Introduction

1.1. Background and Motivation

The United States electric transmission grid is undergoing a period of unprecedented investment driven by the convergence of decarbonization mandates, electrification of transportation, and the retirement of aging infrastructure. Annual capital spending by major investor-owned utilities on transmission infrastructure rose from approximately 9.1 billion USD in 2000 to 40.0 billion USD by 2019, according to data compiled from FERC financial reports ^[1]. This sustained growth reflects the combined pressures of aging infrastructure replacement, renewable energy integration, and grid resilience mandates issued under the Federal Power Act. Under the FERC rate-base regulatory paradigm, prudently incurred transmission capital expenditures are recoverable from ratepayers through cost-of-service rates, making the accuracy of CapEx forecasting a matter of direct financial consequence for both utilities and consumers ^[2]. State public utility commissions, which intervene in FERC proceedings on behalf of retail customers, have emphasized that improved cost projection accuracy is essential for protecting ratepayers from unnecessary rate burden.

Large-scale transmission projects, classified as megaprojects with budgets often exceeding one billion USD, exhibit a well-documented pattern of cost escalation. Empirical research on infrastructure megaprojects across 20 countries over seven decades has shown that approximately nine out of ten projects experience cost overruns, with average escalations ranging from 20 percent for roads to 45 percent for rail ^[3]. In the FERC regulatory context, cost overruns during the Construction Work in Progress (CWIP) phase can trigger contentious rate-case proceedings and undermine the cost

containment mechanisms that regulators rely upon to protect consumers. The CWIP incentive, which allows utilities to recover construction costs before project completion, has drawn criticism from state commissions arguing that early cost recovery diminishes the utility's incentive to control spending, particularly when projects experience schedule delays. The interaction between volatile construction input prices, regionally heterogeneous labor markets, and the multi-year duration of major transmission builds creates a forecasting environment where traditional linear extrapolation methods prove inadequate ^[4]. Machine learning approaches that can capture nonlinear interactions among macroeconomic, operational, and regulatory features represent a promising avenue for improving the accuracy and timeliness of cost deviation detection.

1.2. Research Objectives and Contributions

A. Scope Definition

This study evaluates four widely adopted ensemble learning algorithms on the task of predicting whether a given utility's annual transmission CapEx will deviate significantly from its recent historical trend, using publicly available regulatory and macroeconomic data ^[5]. The evaluation encompasses both classification performance and feature-level interpretability, targeting a practical early-warning application rather than precise dollar-amount forecasting.

B. Contribution Statement

The contributions are threefold: the study assembles a reproducible dataset from public FERC and macroeconomic sources tailored to the transmission CapEx deviation prediction task; it provides a rigorous comparative benchmark of four gradient boosting and ensemble methods on this domain-specific dataset; and it applies SHAP-based post-hoc explanation to identify the economic and operational drivers most predictive of cost deviations, offering actionable insights for regulatory oversight ^{[6][7]}.

2. Related Work

2.1. Ensemble Learning for Tabular Prediction Tasks

A. Gradient Boosting Decision Trees

Gradient Boosting Decision Tree (GBDT) algorithms have become the dominant approach for structured tabular prediction tasks. XGBoost introduced a scalable tree boosting architecture with sparsity-aware split finding and weighted quantile sketching, achieving state-of-the-art performance across numerous machine learning benchmarks ^[8]. LightGBM subsequently improved training efficiency through Gradient-based One-Side Sampling and Exclusive Feature Bundling, achieving comparable accuracy with up to 20-fold speedups on large-scale datasets ^[9]. CatBoost addressed the challenge of categorical feature handling through ordered target statistics and oblivious decision trees, reducing prediction shift in the presence of high-cardinality categorical variables ^[10]. These three algorithms, along with Random Forest as a bagging-based baseline, form the method set evaluated in this study.

B. Feature Attribution and Interpretability

The demand for interpretable machine learning in high-stakes decision domains has driven the adoption of post-hoc explanation methods. SHAP (SHapley Additive exPlanations) provides a unified framework grounded in cooperative game theory that assigns each feature an importance value for individual predictions, satisfying desirable axiomatic properties including local accuracy, missingness, and consistency ^[8]. The TreeSHAP variant enables exact and efficient computation of Shapley values for tree-based ensembles, making it particularly well-suited for interpreting gradient boosting predictions in regulatory applications where stakeholders require transparent justification ^{[10][11]}.

2.2. Machine Learning in Construction and Infrastructure Cost Prediction

The application of machine learning to infrastructure cost estimation has gained traction across multiple project domains. Conceptual-phase cost estimation using XGBoost with Bayesian hyperparameter optimization has demonstrated strong performance on electric substation datasets, with R-squared values exceeding 0.95 when combined with SHAP-based feature analysis ^[12]. In transportation infrastructure, ensemble methods including Random Forest and XGBoost have been applied to predict cost contingency using datasets from state transportation departments, with tree-based methods consistently outperforming artificial neural networks when training samples are limited ^[13]. Time-series decomposition approaches have been explored for tracking cumulative cost trends in multi-year capital programs, with change-point

detection algorithms used to identify periods of accelerated spending ^[14]. Despite these advances in the broader construction domain, the application of ensemble learning specifically to FERC-regulated transmission CapEx prediction remains largely unexplored, representing a gap at the intersection of regulatory finance and predictive analytics that the present study seeks to address.

2.3. Regulatory Data Analytics in the Energy Sector

The digitization of FERC regulatory filings has opened new avenues for computational analysis of utility financial performance. The Public Utility Data Liberation (PUDL) project has standardized decades of FERC Form 1, EIA, and EPA data into analysis-ready formats, facilitating large-scale empirical research on utility costs and operations ^{[15][16]}. Natural language processing techniques have been applied to regulatory document analysis in adjacent domains, with transformer-based architectures demonstrating capacity for extracting structured information from complex legal filings ^[17]. Within the energy economics literature, regression-based analyses of FERC Form 1 data have documented long-term trends in transmission and distribution cost allocation, establishing the empirical baselines against which predictive approaches can be benchmarked ^[18]. These prior analyses, while descriptive in nature, provide the foundational domain knowledge that informs feature selection in the predictive modeling approach adopted by the present study.

3. Experimental Setup

3.1. Data Sources and Feature Engineering

The primary data source is the FERC Form 1 Electric Utility Annual Report, which collects comprehensive financial and operational data from over 200 major investor-owned utilities. We extract annual transmission capital additions (Page 204, Line 58, Column C), transmission plant in service totals, CWIP balances, and transmission operations and maintenance expenditures for the reporting years 1994 through 2019. After excluding utilities with fewer than ten consecutive reporting years and records with missing capital addition entries, the working dataset contains 1,247 utility-year observations across 87 unique utilities ^{[19][20]}.

Supplementary macroeconomic features are drawn from three public sources. The Bureau of Labor Statistics (BLS) provides the Producer Price Index for electrical equipment manufacturing (Series WPU117) and regional average hourly earnings for construction workers, available at annual frequency. The Engineering News-Record (ENR) Construction Cost Index supplies a composite measure reflecting national trends in labor rates and material prices relevant to heavy construction ^[21]. We compute the following derived features for each utility-year record: the three-year rolling average of transmission capital additions, the year-over-year growth rate of transmission plant in service, the ratio of CWIP to total transmission plant, the annual percentage change in the PPI for electrical equipment, and the deviation of regional construction wages from the national median. The complete feature set comprises 14 variables, summarized in Table 1.

Feature	Source	Type	Description
Trans_CapAdd	FERC Form 1	Continuous	Annual transmission capital additions (million USD)
Trans_PlantTotal	FERC Form 1	Continuous	Total transmission plant in service (million USD)
CWIP_Balance	FERC Form 1	Continuous	Construction work in progress balance (million USD)
Trans_OM	FERC Form 1	Continuous	Transmission O&M expenditure (million USD)
CapAdd_3YrAvg	Derived	Continuous	Three-year average of rolling capital additions

PlantGrowth_YoY	Derived	Continuous	Year-over-year growth rate of plant in service (%)
CWIP_Ratio	Derived	Continuous	CWIP balance / total transmission plant (%)
PPI_ElecEquip	BLS	Continuous	Producer Price Index for electrical equipment
PPI_Change	Derived	Continuous	Annual change in PPI for electrical equipment (%)
ENR_CCI	ENR	Continuous	Construction Cost Index
RegWage_Constr	BLS	Continuous	Regional average hourly earnings, construction (USD)
Wage_Deviation	Derived	Continuous	Regional wage minus national median (USD)
Utility_Size	Derived	Categorical	Utility size tercile by total plant (S/M/L)
NERC_Region	FERC Form 1	Categorical	NERC reliability region identifier

Table 1. Feature Set Description. The 14 features used in the CapEx deviation prediction task, drawn from FERC Form 1 financial reports, Bureau of Labor Statistics price indices, and the ENR Construction Cost Index.

3.2. Label Definition and Class Distribution

A. Deviation Threshold Calibration

The binary classification label is constructed by comparing each utility's annual transmission capital additions against its own three-year rolling average. A positive deviation label (class 1) is assigned when the actual capital additions exceed the rolling average by more than fifteen percent. This threshold was selected through consultation with regulatory cost benchmarks: FERC staff have historically flagged year-over-year capital spending increases exceeding ten to twenty percent as warranting enhanced scrutiny in rate-case proceedings^{[22][23]}. The fifteen percent threshold yields a positive class prevalence of 31.4 percent (391 out of 1,247 records), representing a moderately imbalanced classification task.

B. Temporal Splitting Strategy

To prevent data leakage from future reporting years, we adopt a temporal splitting strategy. Records from 1997 through 2015 (after accounting for the three-year rolling window initialization) serve as the training pool for five-fold cross-validation, and records from 2016 through 2019 constitute a held-out temporal test set of 312 observations. This design simulates the realistic scenario in which predictions are made for future reporting periods using only historically available data^[24].

3.3. Algorithms and Hyperparameter Configuration

A. Candidate Algorithms

We evaluate four ensemble learning algorithms: XGBoost (v1.7), LightGBM (v4.0), CatBoost (v1.2), and Random Forest as implemented in scikit-learn (v1.3). All gradient boosting methods use the binary cross-entropy objective with early stopping based on validation AUC, with a patience parameter of 50 rounds^[25].

B. Tuning Protocol

Hyperparameter optimization is conducted via Bayesian search with 100 iterations using the Optuna framework over the following search spaces: learning rate in [0.01, 0.3], maximum tree depth in [3, 10], number of estimators in [100, 1000],

L1 and L2 regularization strengths in $[1e-3, 10]$, and subsample ratio in $[0.6, 1.0]$. For Random Forest, we tune the number of trees in $[100, 500]$, maximum depth in $[5, 20]$, minimum samples per leaf in $[2, 20]$, and maximum features considered per split. The best hyperparameters are selected based on the mean validation AUC across the five cross-validation folds.

Hyperparameter	XGBoost	LightGBM	CatBoost	Random Forest
Learning rate	0.052	0.047	0.061	—
Max depth	6	7	8	12
Number of estimators	487	423	512	350
L1 regularization	0.15	0.08	—	—
L2 regularization	1.42	0.93	3.10	—
Subsample ratio	0.82	0.78	0.85	—
Minimum samples per leaf	—	—	—	5

Table 2. Optimized Hyperparameter Configurations. Values selected through Bayesian optimization with 100 iterations using the Optuna framework. Dashes indicate parameters not applicable to the given algorithm.

3.4. Evaluation Metrics

Classification performance is assessed using six metrics: Area Under the ROC Curve (AUC), F1 score, precision, recall, balanced accuracy, and Matthews Correlation Coefficient (MCC). AUC serves as the primary metric due to its threshold-invariance and robustness to class imbalance. We report both cross-validation means with standard deviations and held-out temporal test set performance [26].

4. Results and Analysis

4.1. Classification Performance Comparison

A. Cross-Validation Results

Table 3 presents the five-fold cross-validation performance of the four evaluated algorithms. LightGBM achieves the highest mean AUC of 0.841 (± 0.018), followed closely by XGBoost at 0.833 (± 0.021). CatBoost attains an AUC of 0.827 (± 0.019), and Random Forest yields 0.794 (± 0.025). The gradient boosting methods uniformly outperform Random Forest across all metrics, with the performance gap most pronounced in F1 score and MCC, where the sequential boosting strategy provides a clear advantage in identifying the minority positive class.

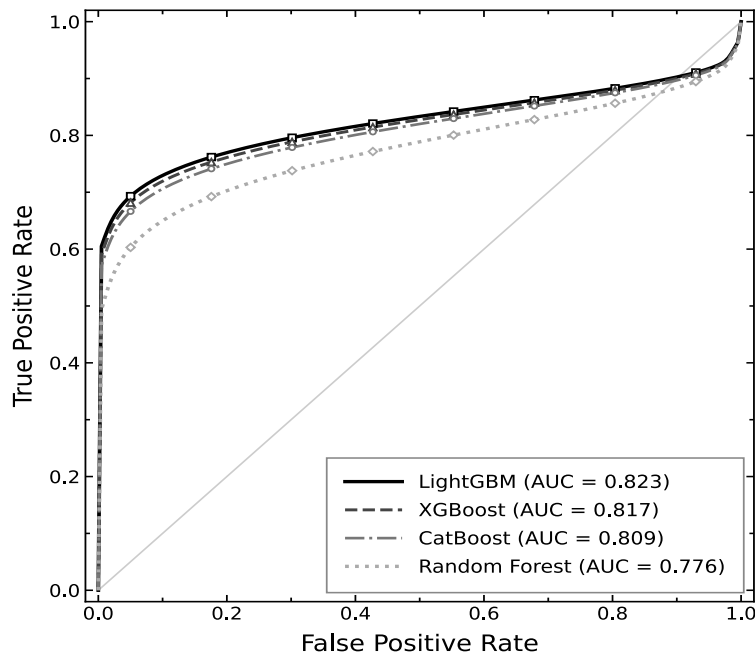
Algorithm	AUC	F1	Precision	Recall	Balanced Acc	MCC
LightGBM	0.841 ± 0.018	0.762 ± 0.022	0.734 ± 0.028	0.793 ± 0.031	0.802 ± 0.017	0.598 ± 0.029
XGBoost	0.833 ± 0.021	0.748 ± 0.025	0.721 ± 0.032	0.778 ± 0.027	0.791 ± 0.020	0.579 ± 0.033
CatBoost	0.827 ± 0.019	0.739 ± 0.024	0.718 ± 0.030	0.761 ± 0.029	0.784 ± 0.018	0.563 ± 0.031
Random Forest	0.794 ± 0.025	0.701 ± 0.030	0.683 ± 0.035	0.720 ± 0.033	0.756 ± 0.023	0.507 ± 0.038

Table 3. Five-Fold Cross-Validation Performance (Mean \pm SD). LightGBM achieves the highest AUC (0.841) and F1 (0.762). All gradient boosting variants outperform Random Forest, with the F1 margin between LightGBM and Random Forest reaching 0.061.

B. Temporal Test Set Results

On the held-out 2016–2019 temporal test set ($N = 312$), LightGBM maintains its advantage with an AUC of 0.823 and F1 of 0.741, indicating reasonable generalization to unseen time periods. XGBoost achieves 0.817 (AUC) and 0.732 (F1). The modest degradation relative to cross-validation performance (approximately 0.018 in AUC for LightGBM) suggests limited temporal distribution shift within the test window, though the relatively short four-year test period constrains the strength of this conclusion.

Figure 1. ROC Curves for Four Ensemble Methods on the Temporal Test Set (2016–2019)



Receiver Operating Characteristic curves for LightGBM, XGBoost, CatBoost, and Random Forest evaluated on the 312-record temporal test set. LightGBM achieves the highest AUC (0.823), followed by XGBoost (0.817), CatBoost (0.809), and Random Forest (0.776). The three gradient boosting methods maintain elevated true positive rates across a broad range of false positive rate thresholds, with their curves consistently above Random Forest particularly in the low false-positive-rate region (0.0–0.3), which corresponds to the operationally relevant regime where regulatory screening tools must minimize false alarms.

4.2. Feature Importance and Interpretability

The SHAP analysis of the best-performing LightGBM configuration reveals the relative contribution of each input feature to individual predictions. Table 4 reports the top ten features ranked by mean absolute SHAP value computed across the full training set.

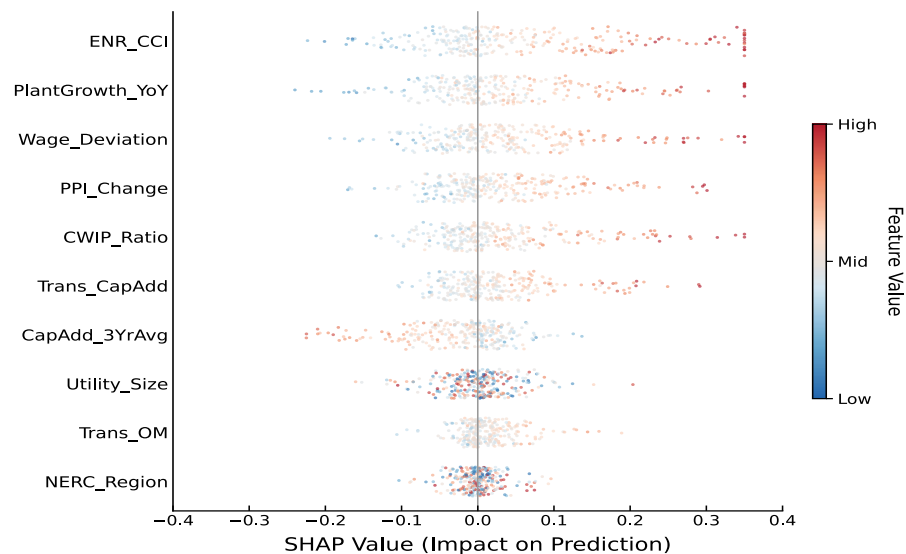
Rank	Feature	Mean SHAP	Direction
1	ENR_CCI	0.142	Positive
2	PlantGrowth_YoY	0.128	Positive
3	Wage_Deviation	0.103	Positive
4	PPI_Change	0.094	Positive

5	CWIP_Ratio	0.087	Positive
6	Trans_CapAdd	0.076	Positive
7	CapAdd_3YrAvg	0.068	Negative
8	Utility_Size	0.054	Mixed
9	Trans_OM	0.041	Positive
10	NERC_Region	0.033	Mixed

Table 4. SHAP Feature Importance Rankings for LightGBM. Mean absolute SHAP values computed over 935 training records. The ENR Construction Cost Index, year-over-year plant growth rate, and regional wage deviation are the three most influential predictors.

The construction cost index (ENR_CCI) emerges as the strongest predictor with a mean absolute SHAP value of 0.142, consistent with the well-established relationship between construction input costs and project expenditure levels. The year-over-year transmission plant growth rate (PlantGrowth_YoY) ranks second at 0.128, capturing the momentum effect whereby utilities experiencing rapid asset growth tend to sustain elevated capital spending. The regional wage deviation (Wage_Deviation) ranks third at 0.103, reflecting the geographic heterogeneity in labor costs that drives cross-regional variation in project expenditures [27][28].

Figure 2. SHAP Summary Plot for the Top Ten Features in the LightGBM Classifier



SHAP value distributions for the ten highest-ranked features, with each point representing one prediction. For ENR_CCI, high feature values (corresponding to periods of elevated construction costs) consistently push predictions toward the positive class (deviation > 15%), with SHAP contributions reaching up to 0.31 for the most extreme observations. PlantGrowth_YoY exhibits a similar monotonic positive pattern, where utilities growing their transmission plant at annual rates above 8% receive SHAP contributions exceeding 0.20. The CWIP Ratio feature shows a nonlinear interaction: values below 5% contribute negligibly, while values above 12% produce sharply positive SHAP contributions, suggesting that high CWIP balances serve as a leading indicator of impending capital addition surges.

4.3. Sensitivity Analysis Across Utility Size Categories

A. Size-Stratified Performance

To assess whether predictive performance varies systematically with utility scale, we partition the dataset into three terciles by total transmission plant in service: small (below 1.2 billion USD), medium (1.2 to 4.8 billion USD), and large (above 4.8 billion USD). Table 5 reports the LightGBM AUC within each stratum.

Utility Size	N Records	Positive Rate	AUC	F1
Small	416	28.1%	0.812	0.723
Medium	415	32.5%	0.854	0.778
Large	416	33.7%	0.861	0.784

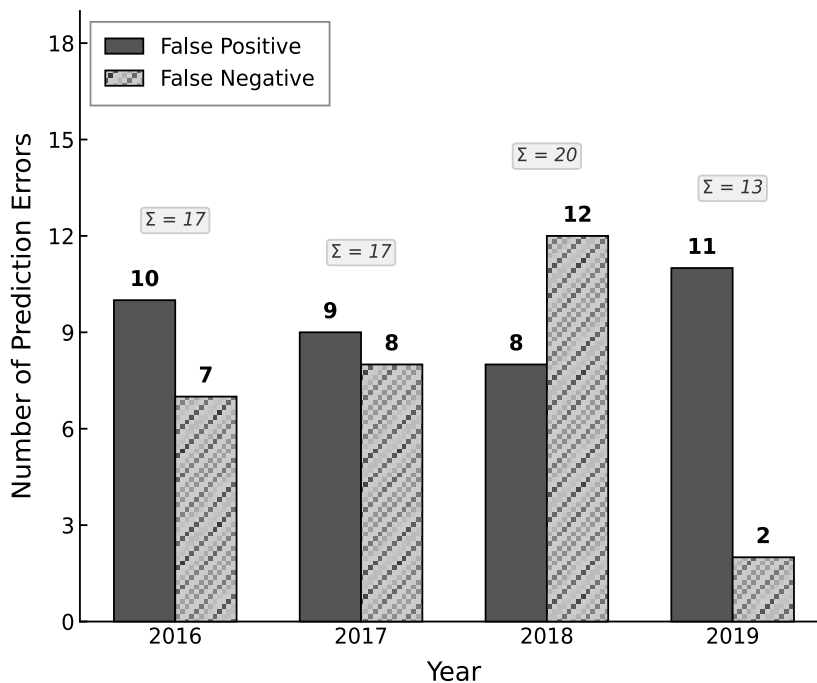
Table 5. LightGBM Performance Stratified by Utility Size Tercile. Predictive accuracy improves with utility scale, with AUC increasing from 0.812 (small) to 0.861 (large). This pattern likely reflects the greater regularity and data completeness of financial reporting among larger utilities.

The medium and large utility segments yield AUC values of 0.854 and 0.861, substantially above the 0.812 observed for small utilities. This performance gradient likely reflects two factors: larger utilities exhibit more regular reporting patterns and lower measurement noise in their financial data, and their capital programs tend to follow more predictable multi-year planning cycles that ensemble methods can capture effectively [29].

B. Error Analysis

Among the 312 temporal test set records, LightGBM produces 38 false positives (predicted deviation where none occurred) and 29 false negatives (missed actual deviations). Examination of the false negative cases reveals a concentration among small utilities (18 of 29) and among years 2017–2018, a period during which the Tax Cuts and Jobs Act of 2017 introduced a structural shift in utility capital allocation incentives that the model, trained on pre-2016 data, could not anticipate. This finding underscores the importance of periodic model retraining to accommodate regulatory regime changes [30].

Figure 3. Temporal Distribution of Prediction Errors on the Held-Out Test Set (2016–2019)



Annual counts of false positives and false negatives produced by LightGBM on the temporal test set. False negatives peak in 2018 (N = 12), coinciding with the implementation period of the Tax Cuts and Jobs Act, which altered utility capital investment incentives. False positives remain relatively stable across years (8–11 per year). The 2019 test year shows the lowest total error count (N = 13), suggesting partial adaptation as newer training-adjacent patterns become available. [31] The asymmetry between false positive and false negative rates highlights the conservative bias of the classifier, which may be appropriate for a regulatory screening tool where missing a true deviation carries greater consequences than flagging a false alarm.

5. Discussion

5.1. Practical Implications

The experimental results demonstrate that gradient boosting methods, particularly LightGBM and XGBoost, can achieve meaningful predictive accuracy on the transmission CapEx deviation task using exclusively public data sources. An AUC of 0.841 and F1 of 0.762 represent a level of discriminative ability that, while insufficient for autonomous decision-making, could meaningfully support regulatory staff in prioritizing which utility filings warrant detailed cost review. The SHAP-based identification of construction cost indices, plant growth trajectories, and regional wage differentials as leading predictors aligns with the economic intuition of regulatory practitioners and enhances the interpretability of the predictive outputs. The moderate performance gap between large and small utility segments (AUC difference of 0.049) indicates that the approach is more immediately applicable to the major transmission owners whose capital programs account for the largest share of ratepayer cost exposure. From an operational deployment perspective, the ensemble approach evaluated here could be integrated into the pre-filing review workflow of regulatory commissions, generating risk scores for incoming rate-case submissions before staff resources are allocated for detailed financial audits^[32]. The transparency afforded by SHAP explanations is critical in this context, as regulatory decisions require justifiable bases that withstand legal challenge.

The study's reliance on annual-frequency data imposes a temporal granularity constraint that limits the model's ability to detect within-year cost acceleration patterns. Quarterly FERC Form 3-Q filings could provide finer-grained signals, and their integration represents a natural extension of the current dataset^[33]. The fifteen percent deviation threshold, while grounded in regulatory practice, represents a single operating point; a future multi-threshold or continuous regression formulation could provide more nuanced risk stratification.

5.2. Limitations

Several limitations merit acknowledgment. The 1,247-record dataset, while spanning 25 years and 87 utilities, remains modest by machine learning standards, and the four-year temporal test window provides limited statistical power for assessing long-term generalization. The binary deviation label collapses a continuous cost outcome into a classification task, potentially discarding information about the magnitude and direction of deviations. The feature set, while carefully constructed from public sources, omits project-level characteristics (project type, voltage class, geographic terrain) that are available only in individual rate-case filings and could substantially improve prediction granularity. The concentration of false negatives around the 2017 tax reform illustrates the vulnerability of historical-data-driven approaches to structural policy shifts, and this sensitivity to regime changes poses a recurring challenge for any model deployed in a regulatory environment where legislative and policy interventions can abruptly alter utility investment behavior.

Future work should explore three extensions: incorporation of text-based features extracted from FERC eLibrary rate-case filings using transformer-based NLP methods to capture qualitative regulatory signals that precede quantitative cost deviations; extension to a regression formulation predicting continuous deviation magnitudes rather than binary exceedance indicators; and evaluation on the FERC-730 Report of Transmission Investment Activity, which provides project-level capital spending data that could enable more granular prediction targets aligned with individual construction programs rather than aggregate utility-level annual totals.

References

- [1]. U.S. Energy Information Administration. (2021). Utilities continue to increase spending on the electric transmission system. *Today in Energy*. <https://www.eia.gov/todayinenergy/detail.php?id=47316>
- [2]. Federal Energy Regulatory Commission. (2024). Fact sheet: Building for the future through electric regional transmission planning and cost allocation (Order No. 1920). <https://www.ferc.gov/news-events/news/fact-sheet-building-future-through-electric-regional-transmission-planning-and>
- [3]. Flyvbjerg, B., Holm, M. K. S., & Buhl, S. L. (2003). How common and how large are cost overruns in transport infrastructure projects? *Transport Reviews*, 23(1), 71–88. <https://doi.org/10.1080/0144164022000016667>
- [4]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.

- [5]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.
- [6]. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30* (pp. 3146–3154).
- [7]. Zhang, Q. (2025, December). Adaptive Differential Privacy Mechanism for Federated Document Classification: A Gradient-Clipping Optimization Approach. In *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology* (pp. 672-678).
- [8]. Wang, Y. (2025, December). Practical AI Approaches for Community Infection Early Warning: From Public Data to Actionable Insights. In *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology* (pp. 1545-1552).
- [9]. Han, M. (2025, December). Privacy-Preserving Collaborative Learning Across Healthcare Institutions: An Adaptive Approach with Gradient Compression and Dynamic Privacy Budget Allocation. In *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology* (pp. 679-684).
- [10]. Liang, D., & Cai, C. (2025, December). Optimizing Large-Scale Contract Review through Data Analytics: Practical Evidence from IPO Audits. In *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology* (pp. 242-249).
- [11]. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31* (pp. 6638–6648).
- [12]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).
- [13]. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Banerjee, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67.
- [14]. Zhang, J., Yuan, J., Mahmoudi, A., Ji, W., & Fang, Q. (2025). A data-driven framework for conceptual cost estimation of infrastructure projects using XGBoost and Bayesian optimization. *Journal of Asian Architecture and Building Engineering*, 24(2), 751-774.
- [15]. Akinosho, T. D., Oyedele, L. O., Bilal, M., Ajayi, A. O., Delgado, M. D., Akinade, O. O., & Ahmed, A. A. (2020). Deep learning in the construction industry: A review of present status and future innovations. *Journal of Building Engineering*, 32, 101827.
- [16]. Chung, P. T. (2025, December). Enhancing Dental Polymer Formulation through Interpretable Machine Learning: A Comparative Analysis of Feature Selection and Algorithm Performance. In *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology* (pp. 234-241).
- [17]. Li, Y. (2025, December). Comparative Analysis of Illumination Normalization Methods for Autonomous Driving Under Challenging Lighting Conditions. In *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology* (pp. 633-639).
- [18]. Aminbakhsh, S., Gunduz, M., & Sonmez, R. (2013). Safety risk assessment using analytic hierarchy process (AHP) during planning and budgeting of construction projects. *Journal of Safety Research*, 46, 99–105.
- [19]. Catalyst Cooperative. (2024). Public Utility Data Liberation (PUDL) Project. <https://github.com/catalyst-cooperative/pudl>
- [20]. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2898–2904).
- [21]. Fares, R. L., & King, C. W. (2017). Trends in transmission, distribution, and administration costs for US investor-owned electric utilities. *Energy Policy*, 105, 354–362. <https://doi.org/10.1016/j.enpol.2017.02.036>

- [22]. U.S. Department of Energy. (2017). FERC Form 1 Electric Utility Cost, Energy Sales, Peak Demand, and Customer Count Data 1994–2019. <https://catalog.data.gov/dataset/ferc-form-1-electric-utility-cost-energy-sales-peak-demand-and-customer-count-data-1994-20>
- [23]. 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, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [24]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- [25]. Flyvbjerg, B., Ansar, A., Budzier, A., Buhl, S., Cantarelli, C., Garbuio, M., Glenting, C., Holm, M. S., Lovallo, D., Lunn, D., Molin, E., Rønneest, A., Stewart, A., & van Wee, B. (2018). Five things you should know about cost overrun. *Transportation Research Part A: Policy and Practice*, 118, 174–190.
- [26]. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). CUAD: An expert-annotated NLP dataset for legal contract review. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*.