PINNACLE ACADEMICS PRESS

Journal of Sustainability, Policy, and Practice

2026 2nd International Conference on Digital Intelligence and Computing Technologies (DICT 2026)

Article

An Empirical Comparison of XGBoost and LightGBM for Capital Expenditure Deviation Prediction in U.S. FERC-Regulated Rate-Base Transmission Projects

Yiyang Peng[1,*] and Shengyan Gao[2]

[1] Master of Public Administration, University of Southern California, Los Angeles, CA, USA
[2] Master of Science in Applied Economics and Econometrics, University of Southern California, Los Angeles, CA, USA
* Correspondence: Yiyang Peng, Master of Public Administration, University of Southern California, Los Angeles, CA, USA

Abstract: U.S. transmission investment is becoming increasingly important as the electric grid expands to support reliability, renewable integration, and long-term regional planning. In a rate-base regulatory framework, deviations between expected and actual transmission CapEx flow to customers through formula-rate true-ups, subject to FERC prudency review. Accurate CapEx deviation prediction therefore has practical value for both FERC prudency review and utility capital planning, yet the application of modern gradient-boosting models to FERC-regulated transmission CapEx data remains underexplored. This paper conducts an empirical comparative evaluation of XGBoost and LightGBM against CatBoost, Random Forest, and an elastic-net baseline, using a utility-year panel assembled from FERC Form 1, EIA-860, FHWA NHCCI, and BLS/FRED macro-commodity series spanning 1994 to 2023. With Bayesian-optimized hyperparameters and a 2016 to 2023 held-out window, XGBoost attains MAE 0.113 and R^2 0.803, marginally ahead of LightGBM at MAE 0.118 and R^2 0.784, while retaining a thirteen-percentage-point advantage over the linear baseline in direction accuracy. Feature-group ablations identify hardware and macro-commodity features as the two largest contributors, with leave-one-out MAE increases of 24.8% and 22.1% respectively. TreeSHAP attributions rank prior-year Construction Work in Progress, copper prices, and Order-1000 era effects among the leading deviation drivers. These findings carry direct practical value for FERC staff reviewing formula-rate filings, utility teams preparing annual CapEx updates, and state commissions intervening in formula-rate proceedings.

Keywords: gradient boosting; capital expenditure prediction; FERC regulation; electric transmission

Received: 28 April 2026
Revised: 17 June 2026
Accepted: 27 June 2026
Published: 02 July 2026

1. Introduction

1.1. Background and Motivation

The reliability of the U.S. bulk power system depends on a multi-decade wave of transmission expansion, as utilities modernize aging infrastructure, connect new generation resources, and respond to evolving regional planning requirements. Prior empirical work using FERC Form 1 data documents rising transmission-related costs among U.S. investor-owned utilities [1], reinforcing the importance of understanding transmission CapEx behavior. Empirical work on U.S. electricity infrastructure finds that transmission overrun risk is institutional in character --- arising from regulatory complexity, multi-party coordination, and mid-project regime transitions rather than from engineering uncertainty alone [2,3]. In a rate-base framework, capital-expenditure deviations affect customer rates through prudency review and formula-rate true-ups administered by the Federal Energy Regulatory Commission (FERC). With transmission capital additions among FERC Form 1 filers reaching approximately twenty-one billion dollars in 2016 and continuing to rise thereafter [4], even moderate utility-year deviations carry dollar-material consequences for ratepayers. Accurate utility-year CapEx deviation prediction therefore has practical value for both FERC staff reviewing transmission filings and utilities preparing annual formula-rate updates, yet modern gradient-boosting ensembles remain underexplored in the context of FERC-regulated transmission CapEx data.

From a forecasting perspective, utility-year CapEx deviation is a tabular, mixed-type, panel-structured prediction task with few strong natural covariates and a heavy regulatory overlay. Modern gradient-boosting ensembles have performed strongly on similar tabular prediction tasks, and the availability of standardized federal datasets makes FERC-regulated transmission an underused empirical testbed. A credible prediction of deviation, paired with auditable per-prediction attributions, could provide FERC staff with a triage signal for prudency review, give utilities an internal sanity check before filing annual updates, and inform rate-design or incentive-mechanism reforms by surfacing systematic deviation drivers. The utility-year panel constructed for this study is itself a public-interest research asset: to our knowledge, it is the first reproducible long-horizon dataset linking FERC Form 1 plant-in-service and CWIP records to interconnection-side and macro-commodity covariates at scale.

1.2. Research Gap, Questions, and Contributions

1.2.1. Research Gap and Three Research Questions

Ensemble learning has been applied extensively to construction and highway cost prediction and, more recently, to electric-substation and industrial project cost estimation, yet the FERC rate-base transmission setting remains underexplored in that literature. In prior work, the authors used FERC Form 1 and macroeconomic indicators to classify whether annual transmission CapEx exceeded a rolling historical benchmark [5]; the present study extends that binary screening framework by modeling continuous utility-year deviation magnitude, expanding the sample through 2023, and evaluating feature-group contributions and TreeSHAP attributions in a regression setting. The institutional features of that setting, including formula-rate true-ups, post-2011 FERC planning reforms, and regulatory cost-review processes, create deviation dynamics that generic construction datasets cannot capture. Three research questions follow. RQ1 asks whether XGBoost or LightGBM delivers better predictive accuracy and direction accuracy for utility-year transmission CapEx deviation using FERC Form 1 data. RQ2 asks which of four feature groups --- hardware, regulatory-era, macro-commodity, and labor --- contributes most to accuracy, and whether those contributions align with sector-specific expectations about transmission CapEx. RQ3 asks whether pre-training on an adjacent public cost-index series, the FHWA National Highway Construction Cost Index, transfers usefully to FERC transmission CapEx deviation prediction --- a question of broader interest because both settings are exposed to common input-price drivers such as steel, aluminum, copper, and construction labor.

1.2.2. Contributions and Paper Roadmap

The paper makes three contributions, each addressing a distinct gap in the prior literature. The first is a data contribution: to our knowledge, this is the first reproducible public-data, multi-decade utility-year panel for FERC-regulated transmission CapEx deviation, integrating FERC Form 1 page-204 transmission plant additions and page-216 CWIP with EIA-860 interconnection-side features, FHWA NHCCI, and BLS/FRED macro-commodity and labor series across 1994--2023. Prior empirical work on FERC transmission economics has relied either on cross-sectional snapshots or on hand-curated project-level samples; neither supports the long-horizon machine-learning analysis that the institutional features of rate-base regulation require. The second is a methodological contribution: building on the authors' prior binary classification framework, this study provides the first regression-based head-to-head benchmark of XGBoost and LightGBM, alongside CatBoost, Random Forest, and an elastic-net baseline, for continuous utility-year CapEx deviation magnitude on FERC-regulated transmission data, with Bayesian-optimized hyperparameters, time-split validation, and bootstrap confidence intervals. It extends the gradient-boosting cost-prediction literature from highway and building construction into a regulated-utility setting whose deviation dynamics are shaped by rate-base recovery and formula-rate true-ups rather than by competitive bidding. The third is an interpretability contribution: a TreeSHAP-based attribution analysis that quantifies the relative contribution of hardware, regulatory-era, macro-commodity, and labor feature blocks, and demonstrates a substantively interpretable interaction between the post-Order-1000 regime and commodity-price cycles.

Section 2 reviews ensemble-learning and cost-prediction literature. Section 3 describes data, feature engineering, training protocol, and evaluation. Section 4 reports aggregate results, hyperparameter and sample-size sensitivities, feature-group ablations, the NHCCI transfer experiment, and SHAP attributions. Section 5 discusses practical implications, limitations, and extensions.

2. Related Work

2.1. Gradient Boosting Foundations and Head-To-Head Benchmarks

2.1.1. From Functional Gradient Descent to Regularized Tree Boosting

Gradient boosting was formulated as stagewise functional gradient descent, introducing the TreeBoost variant that remains the conceptual ancestor of modern implementations [6]. As the canonical bagging baseline in this benchmark, Random Forests delivers provable generalization-error convergence in the number of trees, with the out-of-bag and permutation-importance diagnostics that every subsequent tree ensemble inherits [7]. XGBoost --- with second-order approximations, sparsity-aware split finding, and weighted quantile sketches --- made gradient-boosted trees practical at scale [8]. LightGBM then replaced exact histogram aggregation with Gradient-based One-Side Sampling and Exclusive Feature Bundling, cutting wall-clock training by more than an order of magnitude while preserving accuracy on tabular benchmarks [9]. CatBoost completed the family with ordered boosting and permutation-driven target encoding that removes the prediction-shift bias classical implementations exhibit when many categorical levels appear [10]. An elastic-net penalized regression, which combines L1 and L2 regularization to handle correlated predictors, is included as the linear baseline against which all tree ensembles are benchmarked [11].

2.1.2. Head-To-Head Benchmarks and Hyperparameter Sensitivity

A benchmark of the three modern boosters across UCI-style tabular datasets finds them closely matched on accuracy, with CatBoost slightly ahead on average and LightGBM more sensitive to tree-shape hyperparameters at moderate sample sizes [10,12]. A broader evaluation across 45 tabular datasets further confirms that tree ensembles consistently outperform deep learning methods on structured tabular data with heterogeneous feature scales. The sensitivity pattern motivates joint Bayesian optimization over tree-shape and regularization parameters rather than a fixed grid [13]. These benchmarks, however, draw on cross-sectional datasets that share neither the temporal panel structure nor the regulatory overlay of the FERC setting, motivating a domain-specific comparison on FERC Form 1 data.

2.2. ML for Construction, Infrastructure, and Power-Sector Cost Prediction

A decade of applied work has established ensemble learners as the dominant tabular method for construction cost prediction. XGBoost reaches R^2 near 0.96 on green-building CapEx, with results reported across MAE, RMSE, MAPE, and R^2, and delivers MAPE of roughly 0.31 on archival public-works project data with SHAP-ranked initial-contract amount and scope changes among the leading drivers [14,15]. Pioneering ensemble-classifier work combining text-mining features with numerical descriptors demonstrated that stacked ensembles outperform single learners on overrun classification [16]. Head-to-head comparisons of modified decision trees, LightGBM, and XGBoost on heavy-equipment residual-value regression find LightGBM and XGBoost consistently ahead of the decision tree baseline [17]. Comparisons of multiple regression, artificial neural networks, and XGBoost for early highway-construction cost estimation under owner- and contractor-perspective framings find XGBoost to dominate with substantially smaller training data [18]. None of these studies, however, addresses a rate-base regulated setting. Highway and building construction often operate under fixed-price or unit-price contracting, whereas FERC formula-rate utilities recover prudently incurred costs after the fact via true-ups. This institutional difference matters for both the feature set and the interpretation of attributions, motivating a FERC-specific panel and regulatory-era feature block. Closest to the present setting, the authors' previous work applies XGBoost, LightGBM, CatBoost, and Random Forest to classify whether FERC-jurisdictional utilities experience above-threshold annual transmission CapEx deviations using FERC Form 1 and macroeconomic indicators [5]. That study provides an important binary-screening benchmark, but it does not estimate continuous deviation magnitude or evaluate feature-group contributions in a regression framework.

2.3. FERC Economics and Interpretability Methods

A parallel empirical strand studies transmission CapEx outside the machine-learning frame. Linear-regression analyses of two decades of FERC Form 1 filings have mapped load-side drivers to transmission-and-distribution spending, while hypothesis-testing studies of over four hundred electricity-infrastructure projects have placed transmission lines at the low end of overrun risk and argued that the risk is institutional rather than avoidable [1,2]. A decomposition of fifty years of U.S. nuclear construction cost into sixty-one accounting line items demonstrates that soft non-hardware factors dominate the overrun variance, motivating the regulatory-plus-macro axes adopted here [19]. The unified SHAP framework assigns each feature an additive contribution based on Shapley values from cooperative game theory [20]. Tree-ensemble extensions via exact polynomial-time algorithms then enable the per-prediction attributions and global interaction analyses applied in Section 4 [21,22].

3. Experimental Setup

3.1. Data Sources and Label Construction

Five public sources enter the analysis. FERC Form 1 supplies the primary target and the plant-in-service and CWIP stock covariates, spanning 1994 to 2023 and covering investor-owned Major filers under 18 CFR §101. The Visual FoxPro archive for 1994 to 2020 and the XBRL eCollection feed for 2021 to 2023 are parsed with the community-maintained PUDL schema, in which page-204 yields Total Transmission Plant Additions and page-216 yields the CWIP stock. After dropping holding-company duplicates, excluding utilities reporting fewer than five years, and winsorizing the tails at the first and ninety-ninth percentiles, the analytical panel contains 5,487 utility-year records covering 214 unique utilities; training uses the 1994 to 2015 window (4,213 records) and the held-out window is 2016 to 2023 (1,274 records). EIA-860 Schedule 2 supplies interconnection-side covariates, FHWA NHCCI 2.0 delivers the quarterly national highway cost index from 2003 Q1 onwards [18], and BLS PPI WPU117/WPU1171, FRED PCOPPUSDM/PALUMUSDM, and BLS ECI and NAICS-2371 QCEW complete the macro-commodity and labor blocks. The prediction target is utility-year CapEx deviation, defined as the ratio of actual Form 1 transmission plant additions in year t to the expected additions implied by the year-($t-1$) CWIP roll-forward, minus one. The CWIP roll-forward is a public-data proxy rather than a budget figure: utilities file no ex-ante CapEx forecasts at the Form 1 level, and CWIP itself reflects multi-year project pipelines accumulated across reporting periods. The deviation target is therefore measured against a mechanical accounting expectation, not an internal capital plan. Positive values indicate that actual transmission additions exceed the CWIP-implied expectation, and the empirical distribution is right-skewed with median 0.04, ninetieth-percentile 0.31, and ninety-ninth-percentile 0.87. Table 1 summarizes the panel.

Table 1. Panel Summary Statistics for the FERC Utility-Year Analytical Sample, 1994 to 2023.

Metric	Value
Utility-year records	5,487
Unique utilities	214
Time span	1994–2023 (30 years)
Training window	1994–2015 (4,213 records)
Held-out window	2016–2023 (1,274 records)
Median CapEx deviation	0.04
90th-percentile deviation	0.31
99th-percentile deviation	0.87

Source: FERC Form 1 page-204 line 58 and page-216 (CWIP), parsed via the community-maintained PUDL schema; author's aggregation.

3.2. Feature Engineering Across Hardware, Regulatory, and Macro Axes

3.2.1. Hardware and Project-Scale Features from FERC Form 1 and EIA-860

Twenty-three hardware covariates capture the physical scale of each utility-year observation. Transmission plant in service in year $t-1$, CWIP stock, their one-year first differences, and rolling three-year volatility measures define the stock-and-flow core. Customer count, peak demand, and total energy sales drawn from Form 1 pages 304 and 401 capture load-side scale. Interconnection-side features from EIA-860 count new generator interconnection requests, aggregate interconnecting nameplate capacity, and the average interconnection voltage in each utility's service territory. RTO membership and NERC region are one-hot encoded. Pre-2001 missing values in the interconnection variables are imputed with the utility's within-panel median and flagged with an explicit indicator.

3.2.2. Regulatory-Era, Macro-Commodity, and Labor Features

Eighteen soft-factor covariates span the regulatory-era (4), macro-commodity (9), and labor (5) axes. Three binary era flags mark the post-Order-890 window (from 2008), the post-Order-1000 window (from 2012), and the post-Order-2023 window (from 2023); an incentive-ROE flag identifies utilities receiving fifty, one hundred, or one hundred-fifty basis-point adders. FERC Order 1000, issued in July 2011, restructured regional transmission planning and cost allocation on three axes: mandatory regional and interregional coordination, benefit-based cost allocation for new facilities, and elimination of the incumbent utility's federal right of first refusal on regional projects. The post-Order-1000 indicator is included as a regulatory-era feature to test whether the post-2011 planning environment carries predictive information for utility-year CapEx deviation. The post-Order-2023 flag covers only the final year of the observation window and is included for forward-looking completeness; its within-sample predictive contribution is expected to be limited given the single active year. FERC Order No. 2023, issued July 28, 2023 and effective November 6, 2023, governs generator interconnection procedures; Order No. 2023-A followed in March 2024. Because Order No. 2023 became effective late in the final sample year, the post-Order-2023 flag should be interpreted as an end-of-sample control rather than as a basis for substantive inference. FERC Order No. 1920 and subsequent rehearing orders fall outside the observation window and do not enter the empirical specification; they motivate the study's policy relevance rather than its feature design. The macro-commodity block includes year-on-year PPI WPU1171 change, FRED copper and aluminum returns, and NHCCI annual mean and annual peak-to-trough spread. The labor block aggregates BLS ECI construction wage growth, NAICS-2371 employment growth, the ECI--CPI spread, and two derived tightness measures. Table 2 summarizes the feature groups, and Figure 1 diagrams the schema.

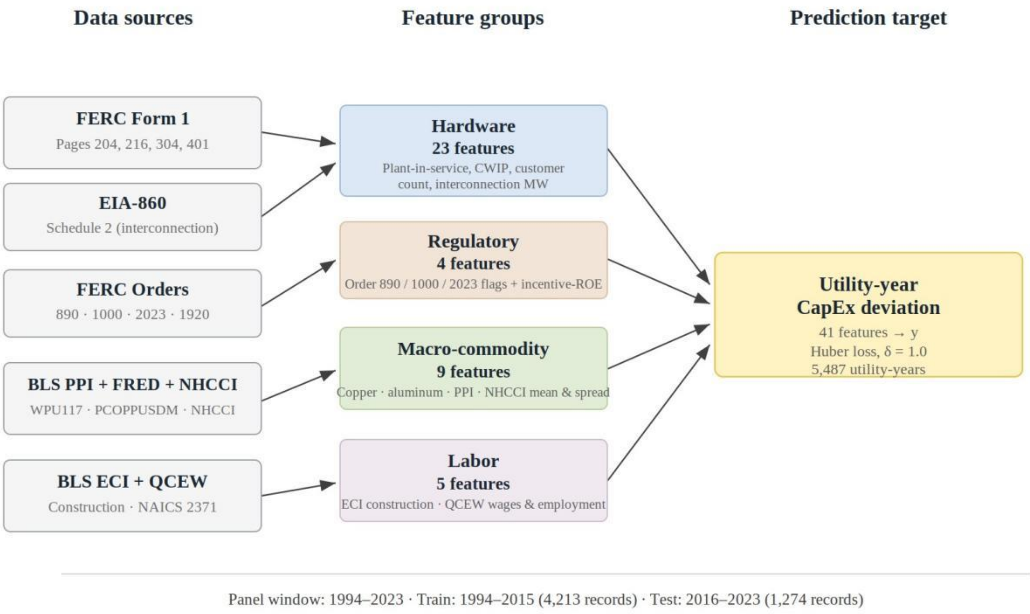

Figure 1. Feature Engineering Schema for the FERC Utility-Year Panel.

Table 2. Feature Groups, Dimensions, and Primary Sources for the 41-Feature Input Vector.

Feature group	Count	Primary sources
Hardware / project scale	23	FERC Form 1 (pp. 204–216, 304, 401); EIA-860 Schedule 2
Regulatory era and incentive-ROE	4	FERC Orders 890, 1000, 2023; eLibrary incentive rulings
Macro-commodity	9	BLS PPI WPU117/WPU1171; FRED PCOPPUSDM and PALUMUSDM; FHWA NHCCI 2.0
Labor	5	BLS ECI construction; BLS QCEW NAICS-2371
Total	41	—

Four feature axes feed the utility-year CapEx deviation target: 23 hardware covariates drawn from FERC Form 1 plant-in-service schedules and EIA-860 interconnection records; 4 regulatory covariates encoding the post-Order-890, post-Order-1000, and post-Order-2023 regimes plus the incentive-ROE adder; 9 macro-commodity covariates drawn from FRED copper and aluminum, BLS PPI WPU117/WPU1171, and FHWA NHCCI; and 5 labor covariates drawn from BLS ECI construction and NAICS-2371 QCEW. Each covariate is aligned to the utility reporting year, and the target is computed from the year-(t − 1) CWIP roll-forward on Form 1 page 216.

3.3. Models, Training Protocol, and Evaluation

3.3.1. Benchmarked Learners and Hyperparameter Optimization

Five learners are benchmarked. XGBoost and LightGBM are the primary comparison; CatBoost, Random Forest, and elastic-net linear regression enter as baselines. Hyperparameters are tuned by Tree-structured Parzen Estimator Bayesian optimization with one hundred trials per learner, five-fold time-series cross-validation that respects the temporal ordering of observations, and early-stopping patience of fifty rounds. For XGBoost the tuning grid covers learning rate, max_depth, min_child_weight, subsample and column subsample, and L1/L2 regularization; for LightGBM the grid substitutes num_leaves for max_depth and leaves the leaf-wise depth unconstrained. The training objective is a Huber loss with $\delta = 1.0$, chosen for robustness to the right-skewed deviation distribution documented in §3.1, where the ninety-ninth-percentile value of 0.87 would exert undue leverage on a squared-error objective. Table 3 documents the selected configurations.

Table 3. Bayesian-optimized Hyperparameter Configurations for Each Learner.

Learner	Key settings
XGBoost	lr = 0.048; max_depth = 6; min_child_weight = 3; subsample = 0.82; colsample_bytree = 0.71; L1 = 0.4; L2 = 1.8; 687 trees
LightGBM	lr = 0.036; num_leaves = 63; min_data_in_leaf = 24; subsample = 0.85; colsample_bytree = 0.68; L1 = 0.6; L2 = 2.1; 452 trees
CatBoost	lr = 0.052; depth = 7; l2_leaf_reg = 3.2; 518 trees
Random Forest	600 trees; max_depth = 18; min_samples_split = 8
Elastic-net	α = 0.008; L1-ratio = 0.38

3.3.2. Evaluation Metrics, Ablations, and Interpretability Pipeline

Primary metrics are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and coefficient of determination (R^2). MAPE is computed across all test-set observations. Given the right-skewed distribution with median actual deviation of 0.04, MAPE figures are sensitive to near-zero denominators and should be interpreted alongside MAE as the primary continuous metric. Direction accuracy, a policy-relevant secondary metric, reports the fraction of test-set observations for which the predicted and realized deviations share a sign, capturing whether a learner correctly flags a utility-year as likely over- or under-spent. Direction accuracy gains are reported relative to the elastic-net. Given the right-skewed target distribution, a majority-class rule that always predicts over-run would itself score non-trivially on this metric, and booster gains should be interpreted with that context in mind. Confidence intervals come from one-thousand-sample bootstrap of the held-out years. Three ablations are run: feature-group leave-one-out; a time-split generalization test with training on 1994 to 2015 and testing on 2016 to 2023; and an NHCCI transfer-learning experiment in which a LightGBM learner is first pre-trained on quarterly highway-cost-index dynamics and then fine-tuned on the utility-year panel. Global interpretability uses the SHAP framework, which assigns each prediction an additive decomposition across features grounded in Shapley-value theory; local case studies use force plots on representative observations.

4. Results and Analysis

4.1. Headline Comparative Results

4.1.1. Aggregate Error and Direction Accuracy

Table 4 reports the five learners' performance on the 2016 to 2023 held-out window. XGBoost attains MAE 0.113 (95% bootstrap CI 0.107--0.119), RMSE 0.165, MAPE 0.263, R^2 0.803, and direction accuracy 0.791, marginally ahead of LightGBM at MAE 0.118 (CI 0.112--0.124), RMSE 0.172, MAPE 0.278, R^2 0.784, and direction accuracy 0.774. CatBoost follows at MAE 0.121 and R^2 0.761, Random Forest at MAE 0.142 and R^2 0.679, and the elastic-net baseline at MAE 0.185 and R^2 0.512. The bootstrap confidence intervals for XGBoost and LightGBM overlap substantially (XGBoost 0.107--0.119, LightGBM 0.112--0.124), and a paired bootstrap of per-observation MAE differences would be needed to formally reject equality; the unpaired CI overlap supports characterizing XGBoost's lead as moderate rather than decisive, consistent with cross-domain evidence that XGBoost retains a marginal advantage over LightGBM across project phases in industrial cost estimation [17]. XGBoost and LightGBM each outpace the elastic-net by thirteen and twelve percentage points respectively in direction accuracy; given the right-skewed deviation distribution, a rule that always predicts over-run would itself score non-trivially on this metric, so booster gains should be interpreted as directional improvement above the elastic-net rather than as an absolute measure of predictive value.

Table 4. Held-out-window (2016--2023) Results for the Five Benchmarked Learners on the 1,274-Record Test Set. Bootstrap 95% Confidence Intervals for MAE Are Reported in Parentheses.

Learner	MAE (95% CI)	RMSE	MAPE	R^2	Dir. Acc.
Elastic-net	0.185 (0.176–0.194)	0.267	0.521	0.512	0.658
Random Forest	0.142 (0.134–0.150)	0.208	0.379	0.679	0.714
CatBoost	0.121 (0.115–0.127)	0.178	0.296	0.761	0.758
LightGBM	0.118 (0.112–0.124)	0.172	0.278	0.784	0.774
XGBoost	0.113 (0.107–0.119)	0.165	0.263	0.803	0.791

4.1.2. Sensitivity to Hyperparameters and Training-Set Size

Figure 2 shows error trajectories across boosting rounds and training-set fractions. LightGBM reaches a plateau near MAE 0.118 after roughly 450 rounds, while XGBoost continues to improve marginally up to 700 rounds before early stopping at MAE 0.113. At small training fractions LightGBM's leaf-wise growth penalizes generalization more sharply than XGBoost's depth-wise alternative: the MAE gap widens to 0.021 at the 10% fraction and narrows to 0.005 at the full-sample condition. Wall-clock training gives LightGBM a roughly 4.8 × advantage per Bayesian-optimization trial in this experimental setup. With smaller training samples, XGBoost is the safer default, and LightGBM requires additional tuning budget on its leaf-count and minimum-leaf-size parameters to match it.

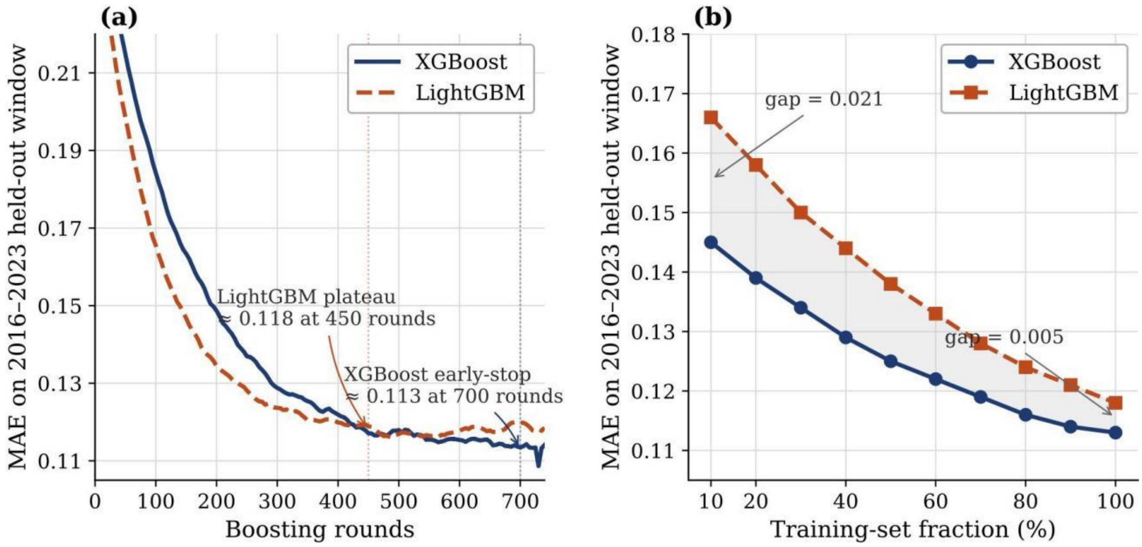

Figure 2. Error Trajectories for XGBoost and LightGBM Across Boosting Rounds and Training-Set Fractions.

(a) MAE on the 2016 to 2023 held-out window as a function of boosting rounds. LightGBM reaches a plateau near MAE 0.118 after roughly 450 rounds, while XGBoost continues to improve marginally up to 700 rounds before early stopping at MAE 0.113. (b) MAE as a function of training-set fraction. The gap between the two learners widens to 0.021 at the 10% fraction and narrows to 0.005 at the full-sample condition, consistent with LightGBM's higher sensitivity to leaf-wise overfitting at moderate sample sizes.

4.2. Feature-Group Ablations and Nhcci Transfer

Table 5 reports the marginal MAE contribution of each feature group via leave-one-out ablation on the XGBoost configuration. Removing hardware features raises MAE from 0.113 to 0.141 (+24.8%); removing macro-commodity features raises it to 0.138 (+22.1%); removing regulatory-era features raises it to 0.132 (+16.8%); removing labor features raises it to 0.124 (+9.7%). Hardware and macro-commodity are the two largest individual contributors, and the regulatory axis, comprising three binary flags and an incentive-ROE dummy, delivers an outsized yield per feature consistent with the soft-factor finding in nuclear construction [19].

Table 5. Feature-group Leave-One-Out Ablation Results on the XGBoost Configuration.

Features removed	MAE	ΔMAE (abs)	ΔMAE (%)
None (full 41-feature model)	0.113	—	—
Hardware (23 features)	0.141	+0.028	+24.8%
Macro-commodity (9 features)	0.138	+0.025	+22.1%
Regulatory (4 features)	0.132	+0.019	+16.8%
Labor (5 features)	0.124	+0.011	+9.7%

The NHCCI transfer-learning experiment reduces MAE from 0.113 to 0.108, a 4.4% improvement consistent with the hypothesis that shared input-price drivers support cross-index transfer, but within the 95% bootstrap confidence interval of the un-pre-trained baseline (0.107--0.119). Pre-training helps most at the start of the test window (2016 to 2018 MAE reduced by 6.8%) and fades as the fine-tuning distribution diverges from the highway-cost pre-training distribution under the post-2022 materials-price regime. The result is best read as suggestive rather than conclusive evidence for cross-index transfer; a multi-index pre-training pool and a paired bootstrap test would be needed to support a stronger claim. A project-level ablation on 312 hand-extracted rate-case true-up observations yields a tighter MAE of 0.091, a 19% improvement over the utility-year headline, consistent with the expectation that utility-year aggregation obscures project-level heterogeneity.

4.3. TreeSHAP Interpretability

4.3.1. Global Feature Attributions

The TreeSHAP beeswarm summary in Figure 3, obtained with a polynomial-time exact-Shapley-value algorithm for tree ensembles [22], ranks the ten highest-impact features by mean absolute SHAP value: prior-year CWIP stock, year-on-year copper price change, lagged transmission plant in service, PPI WPU1171 year-on-year change, BLS ECI construction growth, the post-Order-1000 era flag, customer count, annual NHCCI average, NAICS-2371 employment growth, and the incentive-ROE flag. Stock-and-flow features top the list, commodity and labor features capture the input-price path during construction, and the Order-1000 flag marks a mid-sample regulatory shift in inter-regional planning. The prominence of prior-year CWIP warrants a methodological caveat: CWIP enters the year-($t-1$) denominator of the deviation target by construction, so part of its SHAP magnitude reflects mechanical scaling rather than independent predictive content. The non-CWIP features in the top ten, including lagged transmission plant in service, copper, PPI WPU1171, ECI construction, the Order-1000 flag, customer count, NHCCI, NAICS-2371, and the incentive-ROE flag, are not subject to this caveat and carry the substantive interpretive weight. SHAP interaction values reveal a positive interaction between the Order-1000 flag and copper-price change, plausibly reflecting the long-range regional plans that coincided with the 2011 to 2014 commodity-price peak, and a negative interaction between incentive-ROE adders and CWIP stock, consistent with adders being granted to utilities that have substantially completed their major buildout phases and carry lower incremental CWIP [22].

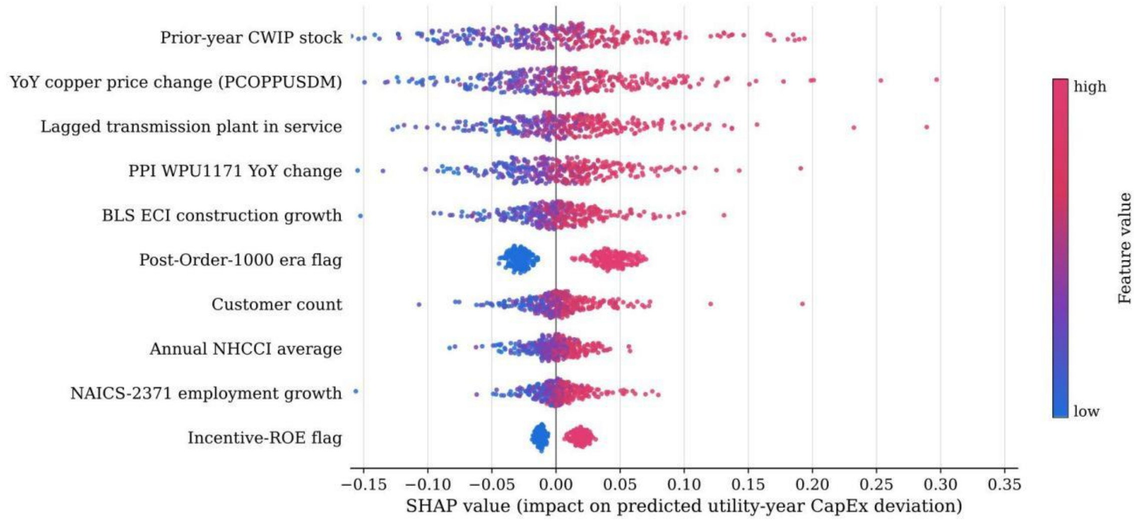

Figure 3. TreeSHAP Beeswarm Summary of the XGBoost Final Fit.

Figure 3 displays the ten features ranked by mean absolute SHAP value from highest to lowest impact, with horizontal position for each observation indicating the magnitude and direction of that feature's contribution to the predicted deviation. Prior-year CWIP stock and year-on-year copper price change display the widest attribution ranges, consistent with their status as the dominant stock-and-flow and commodity-price drivers across the 5,487-record panel.

4.3.2. Local Case Studies

Two utility-year cases illustrate local attribution behavior. In the nominal case, a mid-sized RTO member in 2018 recording a realized deviation of 0.06, SHAP attributions show small positive contributions from copper-price pass-through and the post-Order-1000 flag, offset by a below-average CWIP stock. In the fat-tail case, a large investor-owned utility in 2021 recorded a realized deviation of 0.42, driven primarily by a 38% year-on-year copper-price rise, the post-Order-1000 flag, and an elevated CWIP stock reflecting a multi-year regional transmission plan. Lagged plant-in-service provides a partial offset but does not close the gap. Unlike traditional gain-based rankings, SHAP attributions avoid prediction-shift bias and remain locally accurate [21]. Per-prediction attributions of this form support FERC staff triage under the prudency standard, decomposing the deviation estimate into auditable institutional and market components.

5. Discussion and Future Work

5.1. Summary of Findings and Practical Implications

The study establishes three empirical findings on the first long-horizon public panel of FERC-regulated transmission CapEx deviation. First, modern gradient-boosting ensembles deliver materially better predictive accuracy than the elastic-net baseline traditionally used in transmission cost analysis, with a thirteen-percentage-point gain in direction accuracy that is operationally meaningful for prudency-review triage, though it should be interpreted alongside the right-skewed deviation distribution of the target. Second, XGBoost outperforms LightGBM by a margin that is measurable on aggregate MAE, RMSE, MAPE, and R^2, and more meaningful on direction accuracy (0.791 versus 0.774), yet is not wide enough to license a universal recommendation across rate-base regulatory contexts. Third, the relative contributions of hardware, macro-commodity, and regulatory-era features identified by the ablation and TreeSHAP analyses align with three independent strands of domain reasoning about transmission CapEx, supporting the interpretability of the model rather than treating it as a black box.

The transfer experiment with NHCCI yields a small positive gain that fades as the training and deployment distributions diverge under the post-2022 materials-price regime; transfer from adjacent infrastructure cost indices is worth pursuing with caution rather than as a default assumption. Hardware and macro-commodity features each contribute about a quarter of the MAE budget. The regulatory-era axis, with only three binary flags and an incentive-ROE dummy, delivers an outsized marginal contribution per feature and supports the design choice to encode Federal Power Act regime shifts as explicit covariates. The TreeSHAP attributions align with three domain intuitions: that stock-and-flow variables dominate, that commodity and labor prices matter because the construction window exposes utilities to input-price drift, and that regulatory regime shifts exhibit meaningful interactions with commodity cycles.

Three practitioner implications follow. First, FERC staff reviewing formula-rate annual updates under 18 CFR §35.13 can use the learner's per-prediction attributions as a triage signal that surfaces utility-year filings with atypical deviation drivers, without replacing the existing prudency test, which is a use case that maps directly onto the supervisory workload of the Office of Energy Market Regulation. Second, utility capital-planning teams can use the same attributions as an internal reasonableness check before filing, reducing the incidence of contested true-ups that absorb agency and stakeholder resources. Third, state public utility commissions and consumer-advocate offices intervening in FERC formula-rate proceedings can use the public panel and attribution methodology as a baseline for cross-utility benchmarking, narrowing the information asymmetry that currently disadvantages ratepayer representatives in technical filings. In all three cases, the operational gain is moderate but accrues to processes that govern more than twenty-five billion dollars of annual rate-base recovery [4]. Relative to Random Forest, XGBoost reduces MAE by 0.029 and improves direction accuracy by 7.7 percentage points; relative to elastic-net, it reduces MAE by 0.072 and improves direction accuracy by 13.3 percentage points.

5.2. Limitations and Future Work

Three limitations bound the interpretation of these results. The utility-year level of aggregation obscures project-level heterogeneity that the rate-case true-up ablation partially mitigates but does not resolve; true-up workpapers are heterogeneously structured across utilities, and full-sample extraction was not feasible within the scope of this paper. The panel is restricted to Major investor-owned filers under 18 CFR §101, excluding municipal utilities, electric cooperatives, and federal power marketing agencies whose CapEx recovery mechanisms differ materially from rate-base regulation. SHAP attributions report conditional expectations and not causal effects; external-validity claims about which levers FERC could pull to reduce deviation require a separate identification strategy. Aggregate metrics also obscure tail performance: with a target distribution whose ninety-ninth percentile reaches 0.87, the R^2 and MAE figures reported here are dominated by the bulk of the distribution. The learners' calibration on the right tail, precisely the observations most material for prudency review, is not separately characterized. Reporting test-set residual quantiles, or moving to quantile regression as proposed below, would address this gap directly.

Three extensions merit near-term attention. Project-level extraction from formula-rate true-up workpapers at scale, using the structured Excel schedules already required by existing rules, would allow fine-grained labels without entering the territory of rate-case narrative analysis. Incorporating weather- and outage-driven covariates from FERC Form 714 and from NERC TADS would test whether reliability shocks propagate into CapEx deviation with the expected sign. Replacing the point-estimate regression with quantile regression under a pinball loss would produce calibrated overrun-probability intervals better suited to ratepayer-risk analysis. The contribution of the present paper, a transparent benchmark of five learners on a previously unused public panel, is intended as an enabling step for these extensions.

References

1. R. L. Fares and C. W. King, "Trends in transmission, distribution, and administration costs for U.S. investor-owned electric utilities," *Energy Policy*, vol. 105, pp. 354--362, 2017. [Online]. Available: https://doi.org/10.1016/j.enpol.2017.02.036
2. B. K. Sovacool, A. Gilbert, and D. Nugent, "Risk, innovation, electricity infrastructure and construction cost overruns: Testing six hypotheses," *Energy*, vol. 74, pp. 906--917, 2014. [Online]. Available: https://doi.org/10.1016/j.energy.2014.07.070
3. B. K. Sovacool, D. Nugent, and A. Gilbert, "Construction cost overruns and electricity infrastructure: An unavoidable risk?" *The Electricity Journal*, vol. 27, no. 4, pp. 112--120, 2014. [Online]. Available: https://doi.org/10.1016/j.tej.2014.03.015
4. U.S. Energy Information Administration, "Utilities continue to increase spending on transmission infrastructure," *Today in Energy*, Feb. 9, 2018. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=34892
5. Y. Peng and S. Gao, "Comparative Evaluation of Ensemble Learning Methods for Capital Expenditure Deviation Prediction in FERC Rate-Base Transmission Projects," *Artificial Intelligence and Machine Learning Review*, vol. 6, no. 4, pp. 49--59, 2025.
6. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189--1232, 2001. [Online]. Available: https://doi.org/10.1214/aos/1013203451
7. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5--32, 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324
8. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785--794. [Online]. Available: https://doi.org/10.1145/2939672.2939785
9. G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30*, Curran Associates, 2017, pp. 3146--3154.
10. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems 31*, Curran Associates, 2018, pp. 6638--6648.
11. H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301--320, 2005. [Online]. Available: https://doi.org/10.1111/j.1467-9868.2005.00503.x
12. C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937--1967, 2021. [Online]. Available: https://doi.org/10.1007/s10462-020-09896-5

13. L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why tree-based models still outperform deep learning on tabular data," *Advances in Neural Information Processing Systems*, vol. 35, pp. 507--520, 2022. [Online]. Available: https://doi.org/10.5555/3600270.3600279
14. O. Alshboul, A. Shehadeh, G. Almasabha, and A. S. Almuflih, "Extreme gradient boosting-based machine learning approach for green building cost prediction," *Sustainability*, vol. 14, no. 11, p. 6651, 2022. [Online]. Available: https://doi.org/10.3390/su14116651
15. G. H. Coffie and S. K. F. Cudjoe, "Using Extreme Gradient Boosting (XGBoost) machine learning to predict construction cost overruns," *International Journal of Construction Management*, vol. 24, no. 16, pp. 1742--1750, 2023. [Online]. Available: https://doi.org/10.1080/15623599.2023.2289754
16. T. P. Williams and J. Gong, "Predicting construction cost overruns using text mining, numerical data and ensemble classifiers," *Automation in Construction*, vol. 43, pp. 23--29, 2014. [Online]. Available: https://doi.org/10.1016/j.autcon.2014.02.014
17. A. Shehadeh, O. Alshboul, R. E. Al Mamlook, and O. Hamedat, "Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression," *Automation in Construction*, vol. 129, p. 103827, 2021. [Online]. Available: https://doi.org/10.1016/j.autcon.2021.103827
18. N. Simić et al., "Early highway construction cost estimation: Selection of key cost drivers," *Sustainability*, vol. 15, no. 6, p. 5584, 2023. [Online]. Available: https://doi.org/10.3390/su15065584
19. P. Eash-Gates et al., "Sources of cost overrun in nuclear power plant construction call for a new approach to engineering design," *Joule*, vol. 4, no. 11, pp. 2348--2373, 2020. [Online]. Available: https://doi.org/10.1016/j.joule.2020.10.001
20. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, Curran Associates, 2017, pp. 4765--4774.
21. S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56--67, 2020. [Online]. Available: https://doi.org/10.1038/s42256-019-0138-9
22. S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv*, 2018. [Online]. Available: https://arxiv.org/abs/1802.03888