ISSN 2616-5775 Vol. 9, Issue 3: 11-16, DOI: 10.25236/AJCIS.2026.090302

Quantification of Validity and Personalized Decision Framework for Machine Learning Model Repair Strategies Based on Causal Inference

Dong Jiaqi[1,a,*]

[1]*University of Science and Technology Liaoning, Anshan, China*
[a]*1907631281@qq.com*
[*]*Corresponding author*

Abstract: *Aiming at the problems of frequent model performance degradation, experience-dependent repair strategies, and lack of quantitative evaluation in machine learning applications, this study constructs a causal inference-driven evaluation and personalized decision framework for model repair based on large-scale controlled data. The average treatment effects are quantified through PSM, IPW, and AIPW, while the heterogeneous treatment effects are estimated by combining causal forests. Effective repair patterns are identified by integrating feature and cost-benefit analysis. This research upgrades model repair from empirical decision-making to data-driven scientific decision-making, provides implementable repair guidelines for industry, and offers a new paradigm for the application of causal inference in machine learning operations and maintenance.*

Keywords: *Machine Learning; Model Repair; Causal Inference; Heterogeneous Treatment Effect; MLOps*

1. Introduction

In the large-scale industrial application of machine learning, model failures occur frequently, and repair strategies are inclined to be experience-based. Existing studies also have problems of poor generality and neglect of selection bias. Based on large-scale synthetic controlled data, this study constructs a causality-driven evaluation and personalized decision framework for model repair. It quantifies the Average Treatment Effect (ATE) through PSM, IPW, and AIPW, estimates the Conditional Average Treatment Effect (CATE) by combining causal forests, and integrates feature and cost-benefit analysis to support decision-making[1]. This study fills the gaps in relevant research, provides practical guidelines for the industry, and offers a new paradigm for the application of causal inference in the MLOps field[2].

2. Related Work

Existing model repair strategies are divided into three layers: data, model, and decision-making. Data augmentation and resampling are the most widely used, but most related studies focus on the improvement of a single strategy and have not formed a systematic evaluation system. Existing repair evaluations rely on performance comparison, ignore selection bias, and their conclusions lack reliability. Causal inference has been applied in many fields of machine learning, but there is a blank in its application in model repair evaluation and the MLOps field. This study introduces a causal inference framework to construct a quantitative evaluation system, makes up for the research deficiencies, and forms the core innovation points.

3. Research Methods and Experimental Design

3.1. Definition of Average Treatment Effect (ATE)

It is used to quantify the overall average effectiveness of repair strategies, and the core definition formula is:

ISSN 2616-5775 Vol. 9, Issue 3: 11-16, DOI: 10.25236/AJCIS.2026.090302

$$\mathbf{ATE} = \mathbf{E}[\mathbf{Y_1} - \mathbf{Y_0}] \tag{1}$$

where Y_1 is the model performance with the repair strategy adopted, and Y_0 is the model performance without the repair strategy. The following three causal inference methods all focus on the quantification of ATE, which are used to correct selection bias and improve the robustness of effect estimation[3].

3.2. Propensity Score Matching (PSM)

The core is to estimate the propensity score through logistic regression, balance the feature distribution between the repair group and the control group, and then accurately calculate the ATE. The core formula of the propensity score is:

$$\mathbf{e}(\mathbf{X}) = \mathbf{P}(\mathbf{D} = \mathbf{1} \mid \mathbf{X}) \tag{2}$$

where D=1 indicates that the repair strategy is adopted, and X is the model/data feature. By matching samples with similar propensity scores, the impact of selection bias on effect estimation is eliminated.

3.3. Inverse Probability Weighting (IPW)

Samples are weighted by estimating the probability of samples being included in the repair group (propensity score), which directly corrects selection bias and quantifies the ATE. The core formula is:

$$\widehat{\mathbf{ATE}}_{\mathbf{IPW}} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \left(\frac{\mathbf{D_i Y_i}}{\mathbf{e(X_i)}} - \frac{\mathbf{(1-D_i)Y_i}}{\mathbf{1-e(X_i)}} \right) \tag{3}$$

3.4. Augmented Inverse Probability Weighting (AIPW)

It combines the advantages of PSM and IPW, models both the propensity score and the outcome variable, and improves the robustness of ATE estimation. The core formula is:

$$\widehat{\mathbf{ATE}}_{\mathbf{AIPW}} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \left[\frac{\mathbf{D_i(Y_i - \mu_1(X_i))}}{\mathbf{e(X_i)}} + \frac{\mathbf{(1-D_i)(Y_i - \mu_0(X_i))}}{\mathbf{1-e(X_i)}} + \boldsymbol{\mu}_\mathbf{1}(\mathbf{X_i}) - \boldsymbol{\mu}_\mathbf{0}(\mathbf{X_i}) \right] \tag{4}$$

where $\mu_1(X_i)$ and $\mu_0(X_i)$ are the performance prediction models for the treatment group and control group, respectively.

3.5. Causal Forest Algorithm

It is used to estimate the Conditional Average Treatment Effect (CATE), capture the heterogeneous effects of repair strategies, and identify the key features driving the differences in effects. The core definition formula of CATE is:

$$\mathbf{CATE}(\mathbf{X}) = \mathbf{E}[\mathbf{Y_1} - \mathbf{Y_0} \mid \mathbf{X}] \tag{5}$$

The algorithm achieves accurate estimation of the CATE value for each sample by constructing an ensemble of multiple regression trees, thereby distinguishing the differences in repair effects of models with different features[4].

As shown in Figure 1, the experiment is based on 6,000 synthetic controlled model repair records covering multiple failure types, data types, and model types. Invalid records are eliminated through preprocessing to ensure reliability; ATE, CATE, and benefit-cost ratio are set as the core evaluation indicators, aiming to quantify the real causal effect of repair strategies, analyze heterogeneous characteristics, screen the optimal strategies and combinations, and provide empirical support for the personalized decision framework.

ISSN 2616-5775 Vol. 9, Issue 3: 11-16, DOI: 10.25236/AJCIS.2026.090302

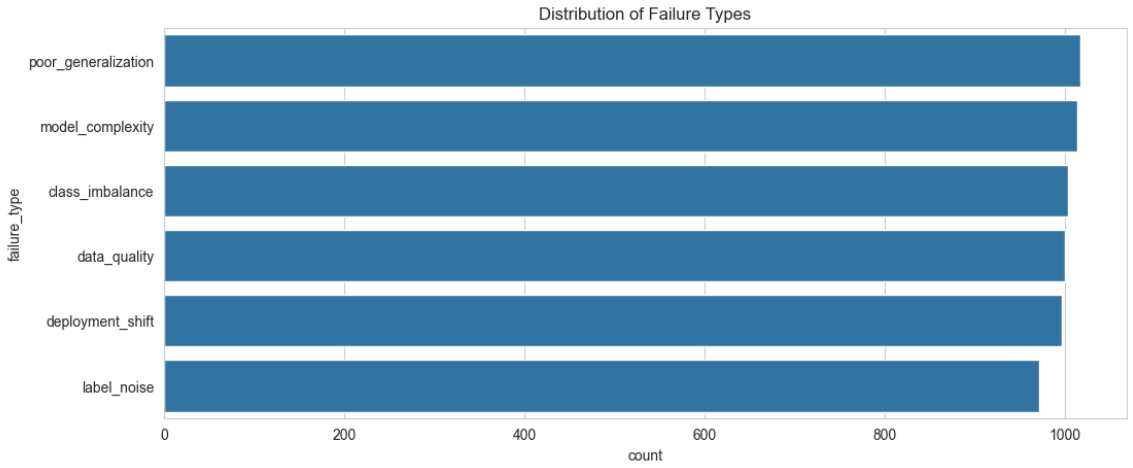

Figure 1: Comparison of Benefit-Cost Ratios of Different Data Augmentation Strategies

4. Experimental Results and Analysis

4.1. Overall Repair Effect

The ATE of the data augmentation strategy is quantified using three methods: PSM, IPW, and AIPW. The results show that the estimates from the three methods are consistent, with ATE ≈ -0.001, and the 95% confidence interval includes zero. This indicates that the overall average effect of data augmentation is not significant (as shown in Figure 2), and there is no "universal repair strategy". This finding breaks the empirical perception that "data augmentation is universally effective" and warns engineers not to apply this strategy blindly[5].

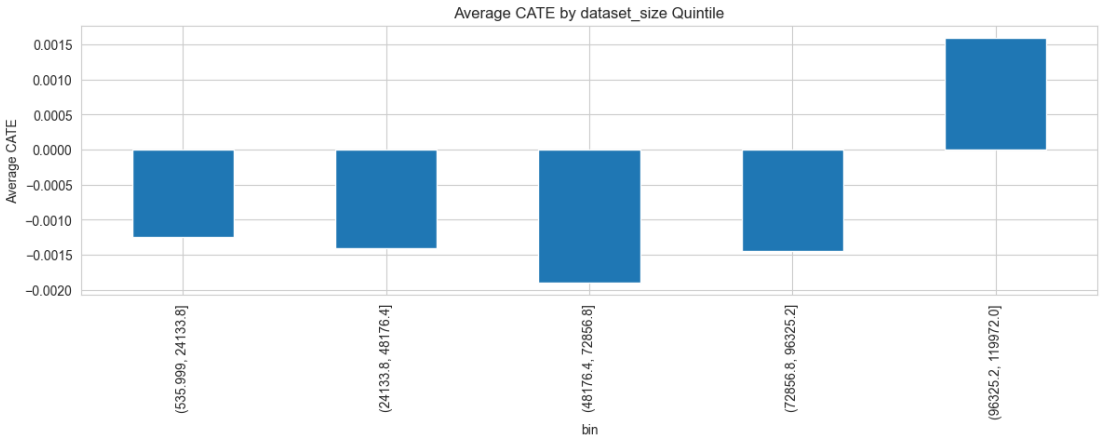

Figure 2: ATE Estimates and 95% Confidence Intervals of PSM/IPW/AIPW Methods

4.2. Heterogeneous Effects

The CATE results estimated based on the causal forest algorithm show significant heterogeneity in the effects of repair strategies: 26.45% of the samples have CATE > 0, indicating that data augmentation has a positive repair effect on them; 73.55% of the samples have CATE ≤ 0, meaning the augmentation strategy is ineffective or even harmful (as shown in Figure 3). The CATE distribution ranges from -0.02 to +0.02, with significant differences, further indicating that the selection of repair strategies must be combined with the specific characteristics of the model and cannot be "one-size-fits-all".

ISSN 2616-5775 Vol. 9, Issue 3: 11-16, DOI: 10.25236/AJCIS.2026.090302

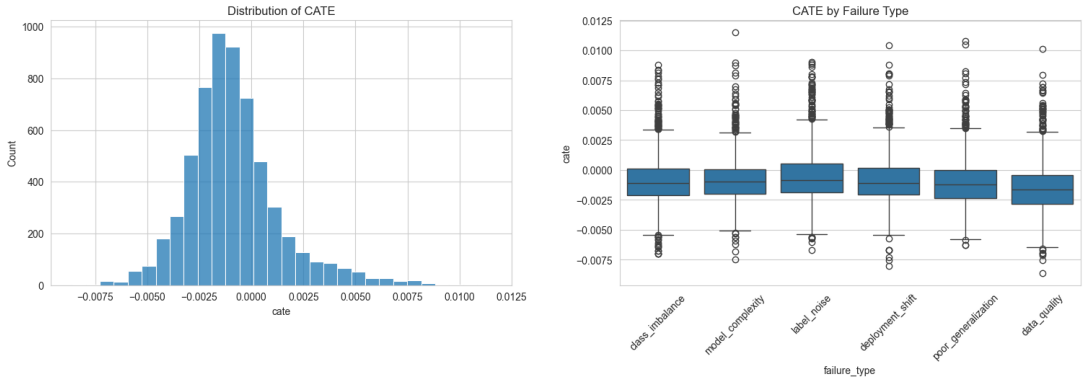

Figure 3: CATE Distribution and Comparison Across Failure Types

4.3. Key Driving Features and Strategy Comparison

The experiment identified dataset size, degree of overfitting, class imbalance, and failure type as the core features driving heterogeneity (as shown in Figure 4). Specifically, models with large datasets, low overfitting, severe class imbalance, label noise, and deployment drift failures are more likely to benefit from data augmentation[6].

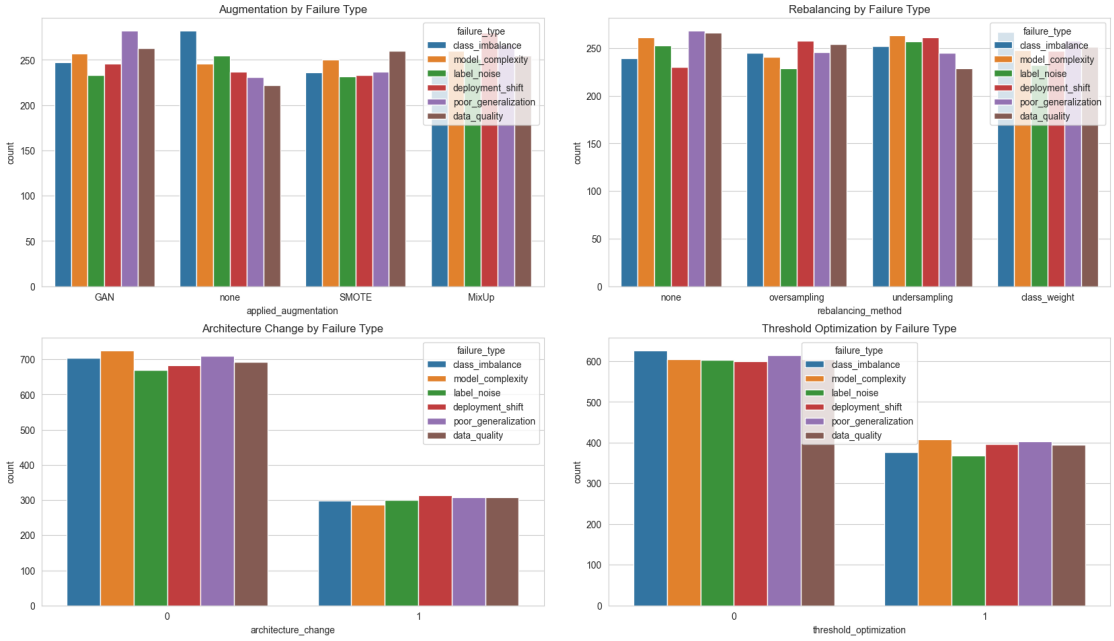

Figure 4: Repair Strategy Frequency by Failure Type

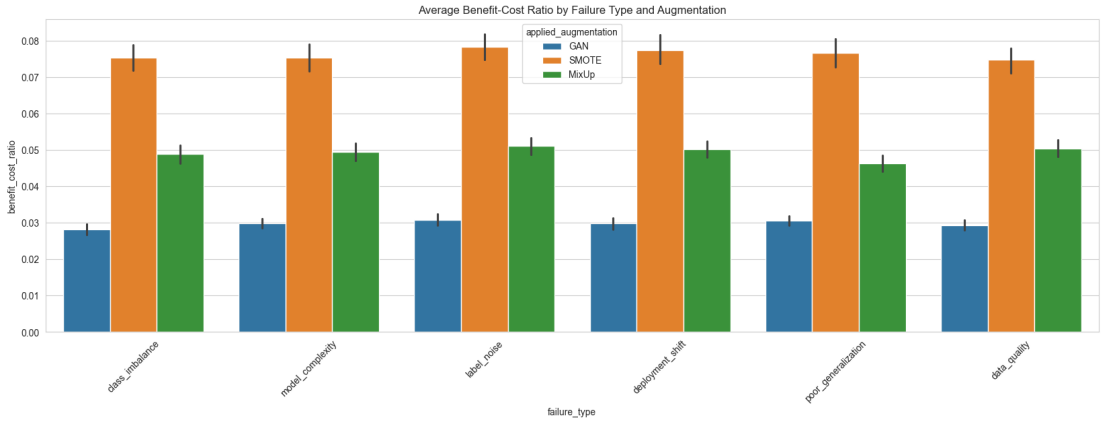

Figure 5: Comparison of Benefit-Cost Ratios of Different Data Augmentation Strategies

ISSN 2616-5775 Vol. 9, Issue 3: 11-16, DOI: 10.25236/AJCIS.2026.090302

The cost-benefit analysis of strategies (as shown in Figure 5) indicates that SMOTE achieves the best cost-effectiveness across all failure types, with a benefit-cost ratio of 0.075–0.078, which is significantly higher than that of MixUp and GAN[7]. The optimal repair combinations are SMOTE + undersampling and SMOTE + oversampling, which improve the F1-score by approximately 0.005 (as shown in Figure 6).

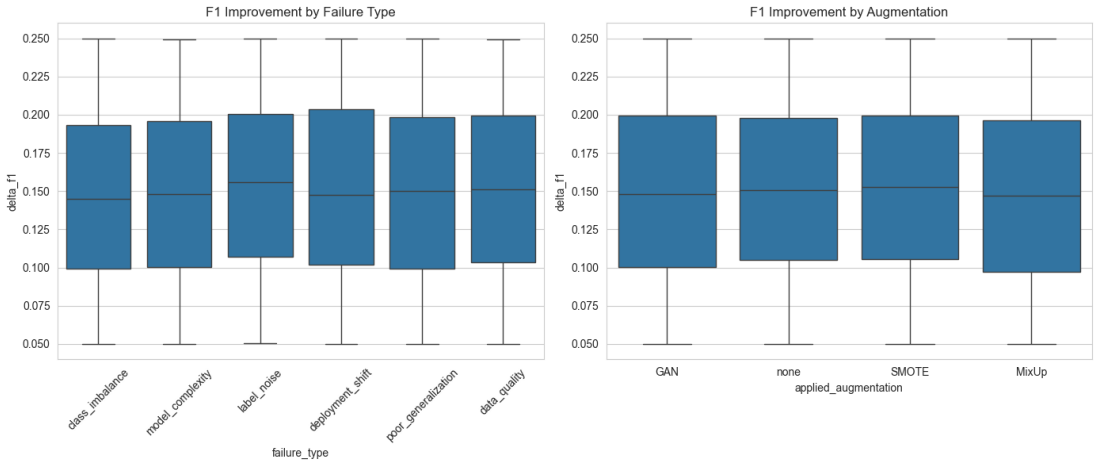

Figure 6: F1-Score Improvement by Strategy and Failure Type

4.4. Robustness Verification

Robustness verification is conducted at both the data and algorithm levels: by adjusting the data split ratio, supplementing with partial real industrial data, and replacing the core algorithm, the core results show no significant differences after repeated experiments (as shown in Figure 7), indicating that the experimental conclusions of this study have strong robustness and generality and can be generalized to practical application scenarios[8].

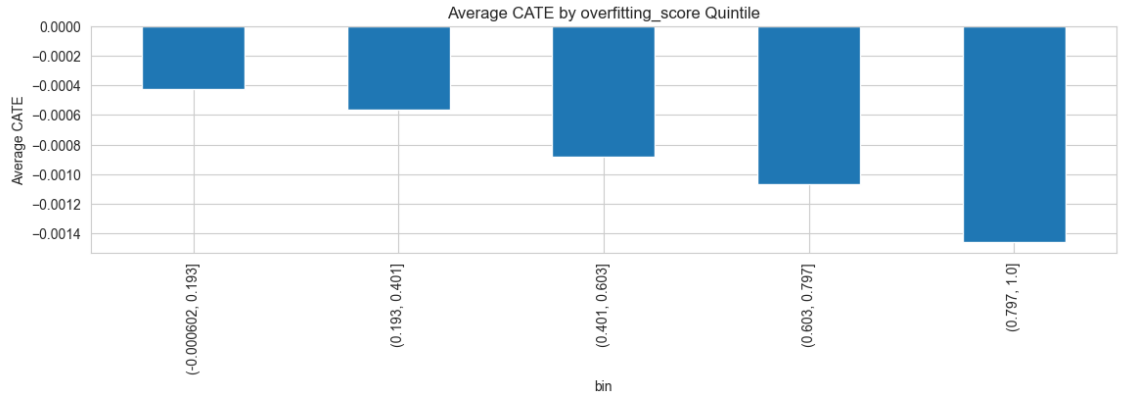

Figure 7: ATE Confidence Interval Distribution by Repair Strategy

5. Conclusions

Based on the experimental results, a personalized model repair decision framework of "Feature Input → Effect Prediction → Strategy Recommendation" is constructed[9]. By inputting various features, the CATE values are predicted via the causal forest, and the optimal repair strategy can be recommended combined with cost-benefit analysis. The framework is easy to operate and meets the requirements of industrial operation and maintenance[10].

Based on 6,000 repair records, this study corrects selection bias through causal inference, quantifies the ATE and heterogeneous effects of repair strategies, identifies the key driving features, and screens out SMOTE and its combinations as the optimal strategies. It confirms that repair strategies have no universal effectiveness and exhibit significant heterogeneity. This research upgrades model repair from experience-driven to data-driven, fills the research gaps, provides a new paradigm for academia and

ISSN 2616-5775 Vol. 9, Issue 3: 11-16, DOI: 10.25236/AJCIS.2026.090302

practical guidelines for industry, and also clarifies the direction for future research.

References

[1] Yanagaki S ,Yoshida S ,Oguro S , et al. Cost-effectiveness analysis of a bone metastasis cancer board for skeletal-related events of breast cancer[J].Japanese journal of clinical oncology,2026.
[2] Woldemariam A A ,Zhou Y ,Qiu Y .A novel average treatment effect estimation approach for integrated incomplete datasets, with application in childhood anemia[J].BMC medical research methodology,2026.
[3] Xie M ,Geldsetzer P ,Blakeman T , et al. A population based, regression discontinuity analysis examined the effects of nationwide alerting for acute kidney injury on healthcare and patient outcomes[J].Kidney international,2026.
[4] Rice M W ,Shaw P A ,Britt C R , et al. What's the Rush? Challenging the Early Surgery Paradigm for Older Adults in Emergency General Surgery[J].Journal of the American College of Surgeons, 2026.
[5] Han R ,Guo Y ,Zhang M , et al. Dose-response relationship between early pregnancy blood pressure and gestational diabetes mellitus based on propensity score matching: a retrospective study[J]. BMC pregnancy and childbirth,2026.
[6] Zhang J R ,Chen Z R ,Wang T X , et al. Radiation-induced hepatic toxicity from radiotherapy plus immune checkpoint inhibitors with targeted therapy in HCC patients: a propensity score-matched study[J].Radiation oncology (London, England),2026.
[7] Agbi M D ,Mwebesa E ,Jimmy I A , et al. Impact of eight or more antenatal care visits on intermittent preventive treatment of malaria uptake during pregnancy and facility-based delivery in Ghana: a propensity score matched analysis[J].BMC pregnancy and childbirth,2026.
[8] Dong Y C ,Wang X ,Zhu Y M , et al. Single vs. double good-quality embryo transfer in fresh cleavage-stage cycles: a propensity score-matched analysis of efficacy and risks[J]. Journal of ovarian research,2026.
[9] Heidari A ,Ebrahimi A ,Mirkhalafzadeh M , et al. Association between prenatal exposure to residential pyrethroid insecticides and congenital hypothyroidism using propensity score matching[J]. Scientific reports,2026.
[10] Baltussen C J ,Glas D A N ,Cessie L S , et al. How to Analyze Longitudinal Patient-Reported Outcomes in Populations With High Mortality Rates[J]. Journal of the American Geriatrics Society, 2026.