Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2025, 13(3), 370-387
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

The Mean Treatment Effect Was Estimated Using a Machine-Learning Model: Evidence from the ECLS-K Dataset

Shenshuo Zhang

School of Medical Sciences, College of Medical and Health Sciences, University of Birmingham, Birmingham, UK
Email: sxz340@alumni.bham.ac.uk

How to cite this paper: Zhang, S.S. (2025) The Mean Treatment Effect Was Estimated Using a Machine-Learning Model: Evidence from the ECLS-K Dataset. *Journal of Data Analysis and Information Processing*, **13**, 370-387.
https://doi.org/10.4236/jdaip.2025.133023

Received: July 30, 2025
Accepted: August 25, 2025
Published: August 28, 2025

Abstract

This study investigates the persistent academic impacts of the Head Start program, a federal government-funded early childhood intervention, using data from the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K). Bayesian Additive Regression Trees (BARTs) are the primary methodology used, and average, conditional, and individual-level treatment impacts on children's mathematics achievement are estimated. BART estimates a negative Average Treatment Effect (ATE) of −1.5421 with increasingly larger adverse effects for children with higher Socioeconomic Status (SES), suggesting diminishing marginal returns. This finding demonstrates the strength of BART to detect nonlinear moderation patterns that are evasive to conventional models. It also implies that Head Start and other preschool interventions will yield greater policy returns when targeted at low-SES children, in order to enable more efficient and fair distribution of public funds. For comparison, Causal Forest estimates a larger ATE (−2.4340) and determines SES to be the overarching moderator, while Propensity Score Matching offers a conservative estimate (−1.2606) without considering effect heterogeneity. These findings underscore the utility of BART in estimating subtle, SES-varying effects of Head Start, and suggest the potential value of more targeted intervention strategies guided by adaptive causal inference.

Keywords

Bayesian Additive Regression Trees (BARTs), Causal Inference, Early Childhood Education, Causal Machine Learning, Nonparametric Estimation

1. Introduction

It has increasingly become necessary for education policy and practice to under-

stand the long-term academic effects of early childhood education interventions. Among them, the Head Start program—designed to provide federally funded preschool education to socioeconomically disadvantaged children—has been highly conspicuous. Nevertheless, empirical evidence for its effectiveness is still ambiguous, usually hampered by issues of methodology in ascertaining causality from observational data.

Traditional causal inference techniques, such as Propensity Score Matching (PSM), yield interpretable and transparent benchmarks but are limited by their strict parametric assumptions and inability to capture complex interactions in a flexible way. Similarly, relatively newer tree-based methods such as Causal Forest (CF) are more flexible but frequentist in assumptions and susceptible to being sensitive to local data structure. In contrast, Bayesian Additive Regression Trees (BARTs) offer a fully Bayesian, nonparametric method that is particularly well adapted to capturing nonlinearities, handling high-dimensional covariates, and estimating both average and heterogeneous treatment effects in a unified model.

This study leverages the power of BART to estimate the causal impact of Head Start attendance on children's longitudinal math achievement using rich observational data from the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K). By flexibly modeling counterfactual outcomes for each individual, BART facilitates the estimation of robust Average Treatment Effects (ATEs), conditional effects by socioeconomic status (CATE), as well as the full distribution of Individual Treatment Effects (ITEs). To situate and validate the findings in context, we also compare results from BART with those from PSM and CF, using them as methodological benchmarks. The comparison framework highlights not only BART's modeling flexibility and interpretive insight but also the critical importance of controlling for treatment effect heterogeneity in early intervention research.

2. Literature Review

2.1. Overview of Causal Inference

Causal inference is a powerful tool in policy analysis, social science research, and educational economics, and the estimation of average treatment effects is one of its fundamental aims. The potential outcomes framework is one of the theoretical underpinnings of causal inference that assumes every observational unit possesses two possible outcomes: one in treatment and one in control, but only one of them is observed. This causes the basic problem of causal inference—counterfactuals are not observable. To get around this, regular assumptions like unconfoundedness, overlap, and the Stable Unit Treatment Value Assumption (SUTVA) are generally called upon to achieve identification and estimation of causal effects. For simple causal inference problems, there are typically two solutions: statistical solution and scientific solution [1]. Classic statistical approaches to causal inference primarily consist of propensity score matching (by modeling the propensity score of a person to be treated and matching similar persons for causal effect estimation [2]), inverse

probability weighting (weighting individuals by propensity scores to balance the covariate distribution between control and treatment groups [2]), and dual robust estimation (combining the propensity score model and outcome model, and estimating the treatment effect via propensity score). Although traditional methods improve the robustness of ATE estimation [3], they face substantial limitations when dealing with high-dimensional and nonlinear data environments. The Average Treatment Effect (ATE), defined as the expected difference in potential outcomes between treated and untreated states, is given by the expression:

$$\mathrm{ATE} = \mathbb{E}\left[Y(1) - Y(0)\right]$$

where $Y(1)$ denotes the potential outcome if treated and $Y(0)$ the potential outcome if untreated. ATE represents the average causal effect of a binary treatment on a population. In recent years, the machine learning method Bayesian additive regression tree has been widely used in the field of causal inference, which provides a new tool for more accurate estimation of ATE [4].

2.2. Bayesian Additive Regression Trees and Comparative Methods for Causal Effect Estimation

Bayesian Additive Regression Trees (BARTs), proposed by Chipman *et al.* in 2010, represent a flexible nonparametric Bayesian framework for modeling complex functions using a sum-of-trees structure [5]. BART is a Bayesian nonparametric approach founded on additive decision tree modeling, with extensive applications in regression, classification, and causal inference. BART models via the additive sum of numerous trees. To formally represent the structure of BART, the outcome Y_i for unit i is modeled as the sum over m regression trees:

$$Y_i = \sum_{j=1}^{M} g\left(x_i; T_j, M_j\right) + \varepsilon_i$$

In this formulation, $g(x_i; T_j, M_j)$ denotes the prediction from the j-th tree based on its structure T_j and associated terminal node parameters M_j, while $\varepsilon_i \sim N(0, \sigma^2)$ captures the residual error. The overall prediction is the additive combination of all m shallow regression trees, where each tree contributes a small, regularized component to the final model. This ensemble framework enables BART to flexibly model complex nonlinearities and interactions while preserving interpretability and providing posterior uncertainty estimates [5]. The model can adaptively select the most predictive covariates to avoid variable selection bias and is free from having to assume linear relationships, which is apt for intricate nonlinear data. The ECLS-K data includes numerous complicated covariates, and the covariates could have complicated nonlinear patterns and interactions that could be difficult for the linear model to successfully pick up such intricate patterns [6]. With the nonparametric regression ability, BART is able to model adaptive nonlinear relationships and automatically choose the most informative variables, thus avoiding the issue of model misidentification [7]. Meanwhile, BART operates within the Bayesian paradigm and employs MCMC sampling to draw parameters and regression functions from

the posterior distribution. The estimation robustness can be enhanced and natural uncertainty quantification can be achieved [8]-[10]. Hill initially applied BART to causal inference in 2011 and found that it outperformed conventional approaches such as Propensity Score Matching (PSM), Inverse Probability Weighting (IPW), and others, particularly in addressing high-dimensional covariates and nonlinear relationships [4]. In recent years, BART has been mainly used to estimate individual treatment effects, study heterogeneity of treatment effects, and ATE [11]. In addition to BART, two comparative methods are introduced in this study to benchmark treatment effect estimation: Causal Forest and Propensity Score Matching (PSM) with outcome regression.

Causal Forest (CF) is a nonparametric ensemble method designed to estimate Conditional Average Treatment Effects (CATEs) by modifying the traditional random forest algorithm. It adapts the tree-splitting criteria to maximize heterogeneity in treatment effects. The estimated treatment effect for a given covariate profile X_i is computed as:

$$\hat{\tau}_{\mathrm{CF}}\left(X_i\right)=\sum_{j\in\mathcal{N}(i)} w_{ij}\left(\frac{Y_j D_j}{\hat{e}\left(X_j\right)}-\frac{Y_j\left(1-D_j\right)}{1-\hat{e}\left(X_j\right)}\right)$$

where $\mathcal{N}\left(i\right)$ is the set of units in the same leaf node as unit i and w_{ij} are forest-derived weights, D_j is the treatment indicator, and $\hat{e}\left(X_j\right)$ is the estimated propensity score [12].

Propensity Score Matching (PSM), on the other hand, represents a traditional approach based on the potential outcomes framework. It begins by estimating the propensity score using logistic regression, and then matches each treated unit to its nearest untreated counterpart. The Average Treatment Effect (ATE) is computed as:

$$\hat{\tau}_{\mathrm{PSM}}=\frac{1}{N_T}\sum_{i\in T}\left(Y_i-Y_{j(i)}\right)$$

where Y_i is the outcome of treated unit i, and $Y_{j(i)}$ is the outcome of the matched control unit. An additional regression adjustment may be applied to the matched sample to improve robustness. Together, these two comparative methods provide benchmarks from both modern causal machine learning and traditional statistical modeling perspectives [2].

2.3. Overview of the ECLS-K Dataset

The Early Childhood Longitudinal Study-Kindergarten Cohort is a national longitudinal survey organized by the National Center for Education Statistics (NCES) that follows kindergartners from school entry (K grade) through academic life. Information on their cognitive, social, economic, family, and school environment is collected, which is mainly used to study children's early educational experiences and their long-term impact on academic and social development [13]-[15]. As a longitudinal dataset, ECLS-K:2011 provides favorable conditions for causal inference. It contains rich covariate information that facilitates the modeling and iden-

tification of complex causal relationships. Further, it has enough data samples that are sufficient for training and testing machine learning approaches. Dynamic evolution of causality is also achievable through the utilization of time series data [16].

3. Methodology

3.1. Modeling Framework

The current research uses the potential outcomes framework to estimate the Average Treatment Effect (ATE) using observational data. Under the assumptions of unconfoundedness, overlap, and consistency, the causal estimand of interest is identified by explicating the conditional expectations of potential outcomes. Instead of relying on parametric forms, a Bayesian nonparametric method is used to model intricate relationships between covariates and outcomes in a flexible manner.

Bayesian Additive Regression Trees (BARTs) are employed as the main estimation technique [5]. In contrast to the usual regression or matching approaches, BART models the outcome surface as a committee of shallow trees learned by Bayesian backfitting and regularization. This enables the model to estimate counterfactual outcomes even in high-dimensional, nonlinear scenarios. The individual-level treatment effects are then approximated by taking the difference between the predicted potential outcomes under treatment and control, and these effects are pooled to obtain ATE [7]. Standard BART implementations typically use 200 trees with a regularization prior that favors shallow trees, where tree depth and shrinkage are controlled by hyperparameters (e.g., $\alpha = 0.95$, $\beta = 2$) to balance model flexibility and overfitting.

3.2. Comparative Methods: Causal Forest and PSM

In addition to BART as the baseline estimation model, two comparison methods—Causal Forest (CF) and Propensity Score Matching (PSM) with outcome regression—are implemented in this study to serve as a reference for treatment effect estimation. Although all three methods attempt to estimate causal effects from observational data, they distinctly vary in statistical paradigm, model assumptions, and inferential power.

BART is a Bayesian nonparametric ensemble method that integrates posterior treatment effect distributions with robustness in modeling nonlinear relationships and quantification of coherence of uncertainty [4] [5] [8]. Causal Forest, by contrast, operates under a frequentist framework and produces Conditional Average Treatment Effects (CATEs) through recursive partitioning and applies asymptotic theory to perform statistical inference [12]. PSM is an old technique that approximates Average Treatment Effects (ATEs) by matching treated and control units on estimated propensity scores, often through logistic regression, and then adjusting for outcomes [2]. PSM does not inherently support heterogeneous treatment effect estimation or uncertainty intervals.

These three methods map to complementary perspectives: BART emphasizes Bayesian inference and adaptability; CF emphasizes adaptive tree-based partitioning for the detection of heterogeneity; and PSM provides an open benchmark according to traditional statistical assumptions. Combining them enables triangulated treatment effect evaluation, ensuring robustness and interpretability in results across methodological platforms.

3.3. Estimation Procedure

For causal effect estimation from observational data, three modeling methods were employed: Bayesian Additive Regression Trees (BARTs), Causal Forest, and Propensity Score Matching (PSM) with outcome regression. The treatment indicator (x) was defined as whether or not the child attended the Head Start program (P1HSEVER), a federally funded preschool program. The outcome measure (y) was the time-average mathematics competence score, computed by the Item Response Theory (IRT) method (avg_MIRT). It reflects the average of math IRT scores across four waves: spring of kindergarten, Grade 1, Grade 3, and Grade 5, capturing long-term academic performance. In order to account for potential confounding, ten pretreatment covariates (a) were employed: child gender (GENDER), race (WKWHITE), socioeconomic status index (WKSESL), public school enrollment (S2KPUPRI), teacher-rated learning approach (apprchT1), household food stamp participation (P1FSTAMP), single-parent family (ONEPARENT), prior preschool enrollment (WKCAREPK), parent health rating of child (P1HSCALE), and incidence of reported child sadness/loneliness (P1SADLON).

For each individual in the sample, two counterfactual outcomes were predicted: one assuming the individual received the treatment $\hat{Y}_i(1)$, and the other assuming they did not $\hat{Y}_i(0)$. These values were generated by setting the treatment indicator accordingly while holding fixed covariates, and feeding the modified inputs into the fitted model. The Individual Treatment Effect (ITE) was then computed as:

$$\hat{\tau}_i = \hat{Y}_i(1) - \hat{Y}_i(0)$$

The Average Treatment Effect (ATE) was obtained by averaging the ITEs across all individuals:

$$\text{ATE} = \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i$$

Where appropriate, conditional Average Treatment Effects (CATEs) were estimated by stratifying the sample based on covariates and repeating the procedure within each subgroup [4]. The Conditional Average Treatment Effect (CATE) was computed by estimating the expected difference in outcomes within specific subgroups defined by covariates:

$$\text{CATE}(X = x) = \mathbb{E}\left[\hat{Y}(1) - \hat{Y}(0) \mid X = x\right]$$

By adopting this multi-model potential outcome framework, including BART,

Causal Forest, and Propensity Score Matching with outcome regression, flexible estimation of treatment effects is enabled while mitigating reliance on parametric assumptions.

3.4. Implementation Details

All models were implemented under Python 3.10.13. Bayesian Additive Regression Trees (BARTs) fit the econml.dml.BART class of Microsoft's EconML library [17]. Causal Forest was executed with the econml.grf.CausalForestDML class, and Propensity Score Matching (PSM) was implemented with a combination of scikit-learn for logistic regression and pymatch for nearest neighbor matching.

Prior to model fitting, missing covariates were replaced with median values, and all the attributes were standardized via StandardScaler from the scikit-learn library. This method was chosen for its robustness to outliers and skewed covariate distributions. Given the low rate of missingness and the limited number of continuous variables, more complex methods such as multiple imputation were deemed unnecessary. For PSM, outcome regression adjustment was performed after matching to address potential residual imbalance in covariates.

All computations were carried out on a 2023 MacBook Pro with an Apple M3 Max chip and 48 GB RAM, running macOS Sequoia version 15.3.2. GPU acceleration was not employed in the model procedure.

4. Results

4.1. Bayesian Additive Regression Trees (BART)

As a means of flexibly estimating the treatment effect of Head Start enrollment on children's long-term mathematics performance, we fitted a two-model BART using the bartpy implementation. We trained this model on a sample of 7362 children with full covariate data and controlled for 10 pretreatment covariates including gender, race, SES, type of school, and early learning measures. In order to make it replicable, we set the random seed to 42.

We trained separate BART models for treated (T = 1) and untreated (T = 0) groups, and predicted counterfactual outcomes for all individuals to obtain Individual Treatment Effect (ITE) estimates. The average Treatment Effect (ATE), calculated as the ITE mean, was −1.5421, which means Head Start participation was associated with a considerable decline in standardized mathematics scores on average. The distribution of ITEs is shown in **Figure 1**, exhibiting a left-skewed unimodal shape with a mean below zero and only a small minority of positive values.

In an attempt to explore treatment effect heterogeneity, ITEs were stratified by socioeconomic Status (SES) using tertile-based divisions based on the WKSESL variable. Conditional Average Treatment Effects (CATEs) in SES subgroups were:

- Low SES: −0.9797;
- Medium SES: −1.4637;
- High SES: −2.1882.

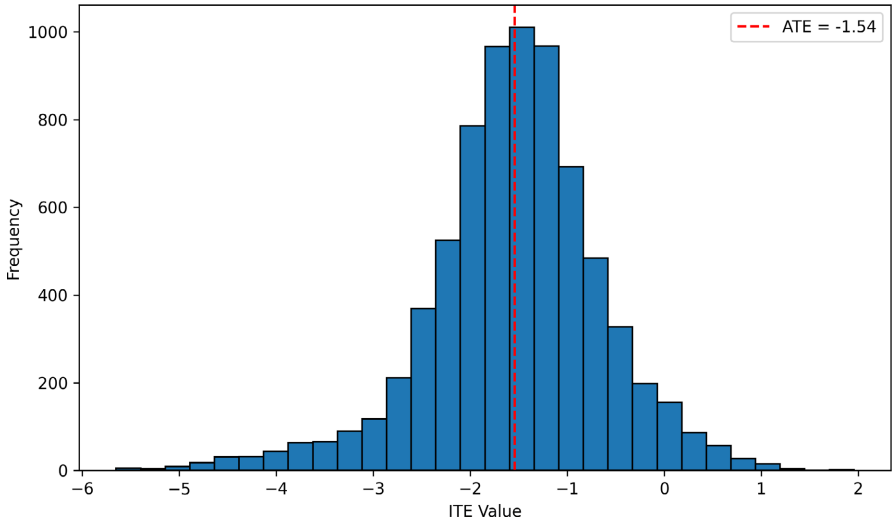

Figure 1. Distribution of individual treatment effects (BART).

These results show a monotonic pattern: children with higher SES origins were more negatively impacted by Head Start attendance. This is consistent with substitution effects or law of diminishing marginal returns hypotheses, where higher SES families enjoy better substitutes or baselines, reducing the relative worth of Head Start. Differences in CATE are well established in **Figure 2**, where effect size rises with level of SES.

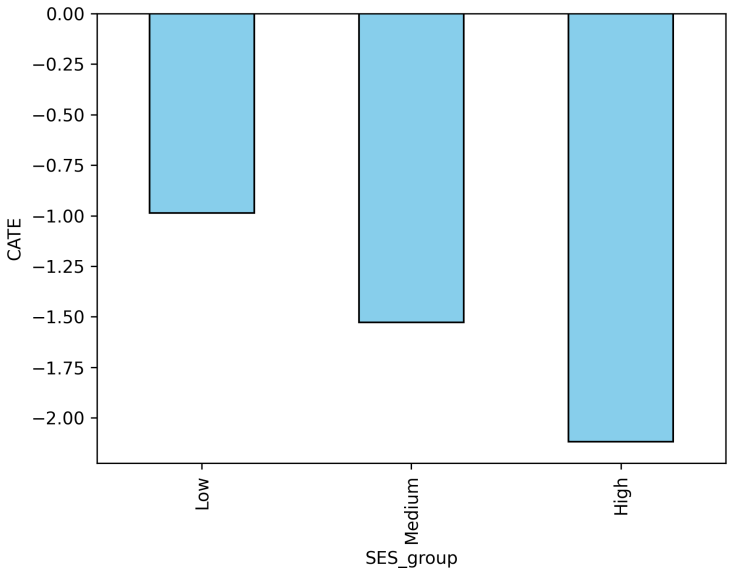

Figure 2. CATE by socioeconomic status group (BART).

In addition, **Figure 3** is a scatterplot of ITEs against the continuous SES variable (WKSESL), which provides additional evidence for a strong negative linear relationship. The fitted regression line shows that as SES increases, the expected individual effect becomes more negative, once again implicating SES as an important moderator of treatment efficacy.

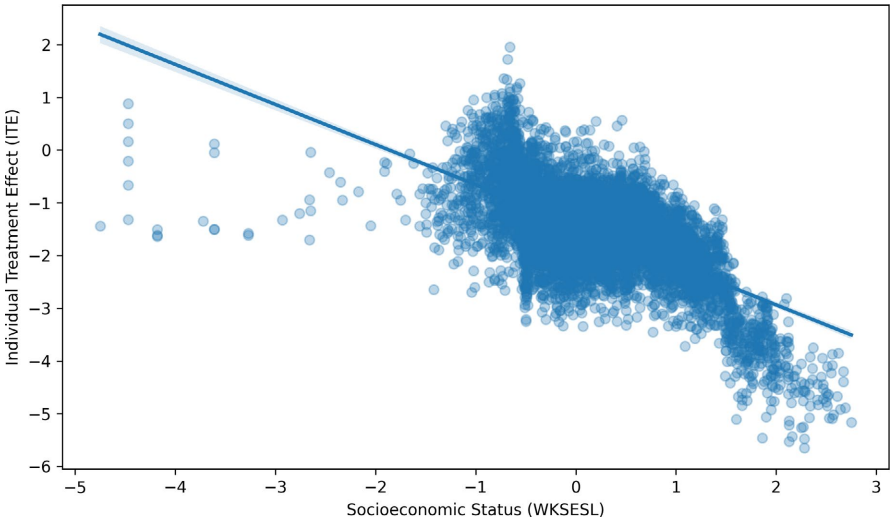

Figure 3. ITE vs. SES (BART).

For full reproducibility, BART output in the form of ITEs, SES scores, and group labels is all saved to bart_ite_individuals.csv. The dataset allows re-generation of the ATE, CATE, and ITE distribution plots without re-running the stochastic BART model.

4.2. Causal Forest

To investigate treatment effect heterogeneity further, we fitted a Generalized Random Forest (GRF) with the econml version of Causal Forest. The model was fit to the same dataset of 7362 observations, with the same 10 pretreatment covariates as the BART model. A total of 500 trees were planted, and honest estimation was disabled to promote stability in small treated subsamples. The random seed was set to 42 for reproducibility.

With regard to its truth-telling counterpart, this model allows for better use of limited treatment group data during tree construction, but at the cost of potentially greater susceptibility to split selection and variance estimation biases.

The estimated ATE was −2.4340, the strongest negative effect of any of the three approaches. This translates to the average of children within Head Start being 2.4 points lower in long-term mathematics outcomes compared to their nontreated counterparts. **Figure 4** provides the full distribution of ITE. As one can easily see from the figure, it is unimodal but very skewed and concentrated at about −2 and −3. Only a few are extreme negative outliers.

To estimate Conditional Average Treatment Effects (CATE), we again stratified the sample into SES tertiles. The resulting group-specific estimates were as follows:

- Low SES: −1.8117;
- Medium SES: −2.7929;
- High SES: −2.7018.

As shown in **Figure 5**, the treatment effect is always strongly negative across all SES groups and is strongest in the medium SES stratum. The pattern somewhat de-

viates from the monotonic gradient in BART and suggests that CF could be more sensitive to local nonlinear interactions between SES and other covariates.

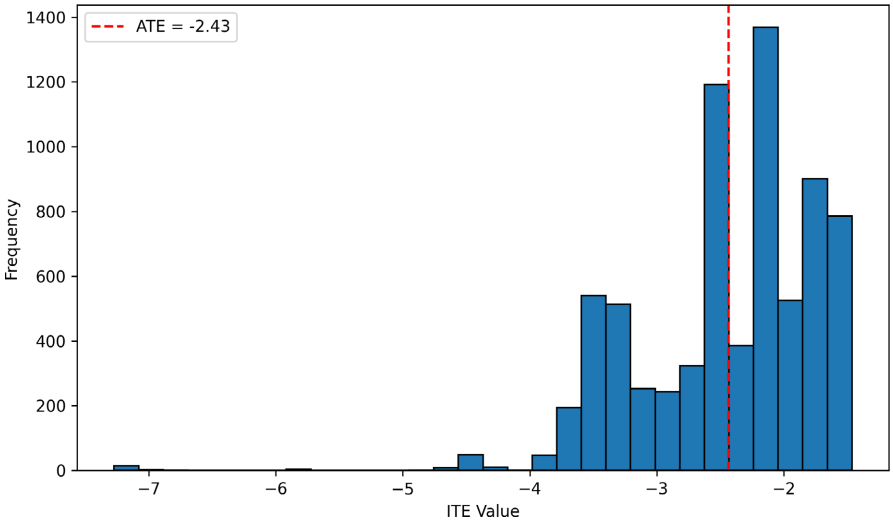

Figure 4. Distribution of individual treatment effects (Causal Forest).

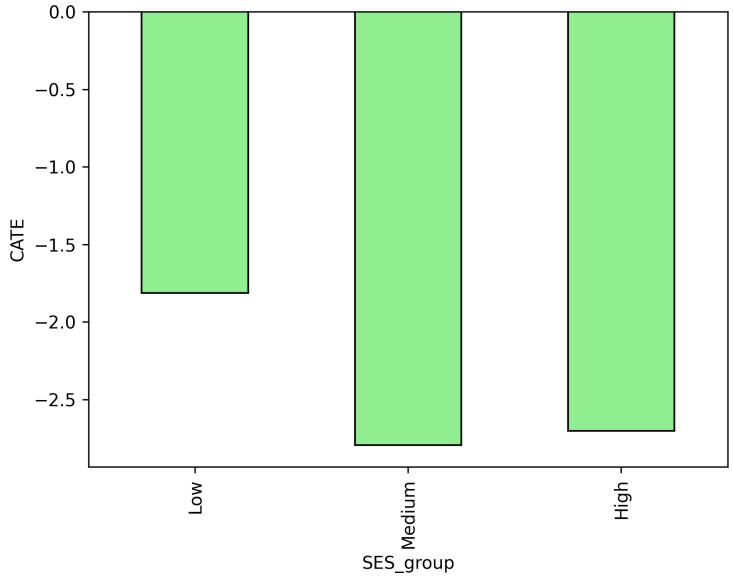

Figure 5. CATE by socioeconomic status group (Causal Forest).

Besides that, we generated feature importances from our Causal Forest model. The results, as indicated in **Figure 6**, show that socioeconomic status (WKSESL) was by far the most important variable, accounting for 66.4% of total importance. Public school enrollment (S2KPUPRI: 27.3%) and food stamp use (P1FSTAMP: 2.5%), which are proxies for family socioeconomic conditions, were the second and third most important variables. The remaining variables were insignificant.

All individual ITE values, SES group assignments, and feature rankings are saved in causal_forest_results_small_sample.csv, enabling complete replication and further subgroup analysis.

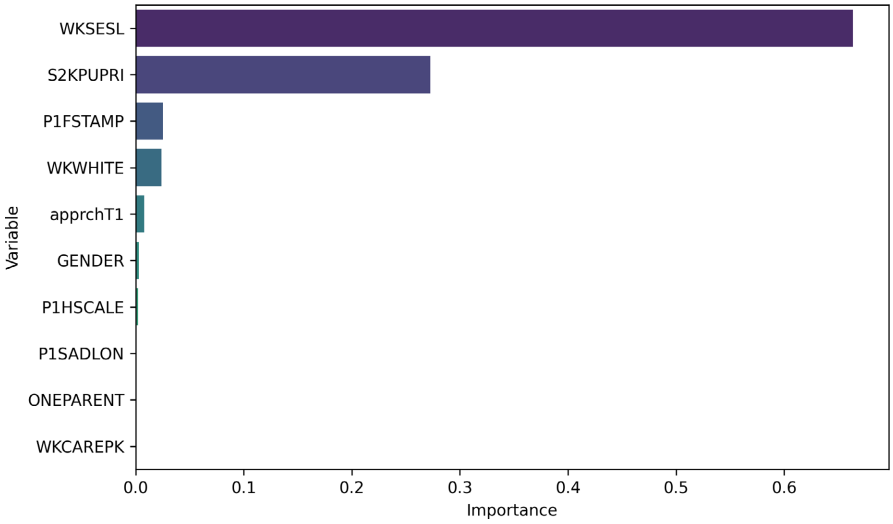

Figure 6. Feature importances (Causal Forest).

4.3. Propensity Score Matching (PSM)

To provide a basis for comparison, we employed Propensity Score Matching (PSM) to calculate the Average Treatment Effect (ATE) of Head Start participation. We predicted the propensity scores with a logistic regression model on the same 10 covariates employed in the BART and Causal Forest models. Due to the over-sampling of treated units and under-sampling of untreated ones (165 treated and 7197 untreated), we performed 1:1 nearest neighbor matching without replacement, yielding a matched sample of 330 children (165 treated and 165 untreated matches).

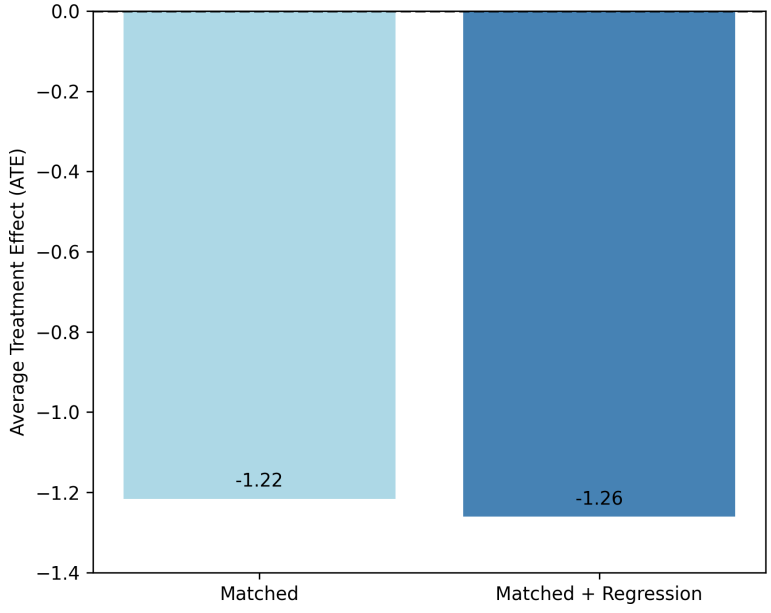

Figure 7. ATE estimates from PSM.

The ATE estimated using the matched sample was −1.2158, indicating that

Head Start participants were, on average, 1.22 points behind their matched controls in mathematics. To further address any residual covariate imbalance in the matched sample, we performed a linear regression of the outcome on the treatment and the covariates. This adjustment resulted in an even stronger ATE of −1.2606.

While PSM cannot be used to estimate Individual Treatment Effects (ITEs) or Conditional Average Treatment Effects (CATE), it is useful to have even a semi-parametric benchmark. As **Figure 7** shows, the regression-adjusted estimate further reinforces the negative effect found with simple matching. Compared to the BART and Causal Forest models, which can handle nonlinearities and uncover effect heterogeneity, PSM provides a conservative lower-bound program impact estimate.

5. Conclusions

This research investigated the causal effect of Head Start attendance on children's math performance based on longitudinal data from the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K). By applying three different causal inference approaches—Propensity Score Matching (PSM), Bayesian Additive Regression Trees (BARTs), and Causal Forests (CFs)—we assessed the average and heterogeneous treatment effects of a large federal early education program. The use of different estimation approaches enabled us to cross-validate results, identify consistent patterns, and assess methodological robustness of causal inference in nonexperimental settings.

The results across all the models revealed the same negative Average Treatment Effect (ATE), and Head Start participation was associated with slightly lower math achievement in subsequent years. While this finding may seem counterintuitive at first blush, it accords with emerging research highlighting the subtlety of long-term impacts of early childhood programs while holding constant fluctuating home environments, school quality, and SES-based opportunity structures [18]. The PSM approach generated a modest ATE estimate of −1.2606, while the flexible machine learning approaches generated bigger effects: BART gave an ATE estimate of −1.5421, and CF generated an ATE estimate of −2.4340. These progressively larger estimates arise because of the greater ability of modern methods to capture subtle interactions and nonlinearities in the data.

Most significantly, both BART and CF revealed significant heterogeneity of treatment effect by Socioeconomic Status (SES). BART showed a unidimensional gradient—better-off children had worse effects—whereas CF showed the largest effects in the middle SES group. These findings suggest that such early interventions as Head Start may not benefit all children equally and indeed may have negative effects among higher-SES participants. Potential explanations include substitution effects, opportunity costs, or differential baseline access to challenging educational opportunities. These findings highlight the importance of targeting and modulating policies to reduce marginal costs for the most disadvantaged chil-

dren.

Methodologically, the study demonstrates the advantage of combining classical and modern causal inference approaches. PSM remains useful as a benchmark estimator with its interpretability and transparency. BART achieves a trade-off between flexibility and interpretability through its regular CATE patterns. Causal Forests, while more complex, excel at capturing fine-grained treatment effect heterogeneity and detecting covariates with the most moderating influence. The triangulated approach not only increases credibility but also enables researchers to view policy effects from different vantage points.

Nevertheless, there are limitations to this study. The number of treated observations was relatively small, and this could affect the stability of subgroup analyses. Residual confounding and measurement error inherent to nonexperimental data may also distort the estimates. Furthermore, we focused exclusively on academic performance; future research should examine socio-emotional or behavioral development as other endpoints. Model extension to include treatment intensity, dosage, or time-varying effects could yield more insights.

Overall, this study highlights that early childhood programs like Head Start may have heterogeneous gains, particularly if not matched with participant background characteristics. Policymakers ought to consider stratified program designs or supplemental supports tailored to particular subgroups. Methodologically, the integration of causal machine learning provides an empirically compelling method for the evaluation of education interventions and warrants more application in policy-relevant social science scholarship.

6. Discussion

This study provides new long-term academic impacts of early childhood interventions by triangulating across three approaches to causal inference. While all specifications indicate a strong negative Average Treatment Effect (ATE) of Head Start attendance on math achievement, this result needs to be interpreted cautiously.

To further interpret these results in light of the methodological frameworks, we synthesized and compared the three approaches on multiple dimensions: estimated ATEs, heterogeneity detection, and model interpretability. This integrated comparison provides clarity on the relative strengths and limitations of each method and enhances the policy relevance of the findings.

To comparatively assess the strengths and limitations of the three causal inference approaches used in this study—Propensity Score Matching (PSM), Bayesian Additive Regression Trees (BARTs), and Causal Forests (CFs)—we examined their corresponding estimates of the Average Treatment Effect (ATE), capacity for detecting treatment effect heterogeneity, and interpretability in terms of variable importance as well as subgroup analysis. While all models were trained on the same dataset and covariates, their estimands and algorithmic assumptions diverged substantially, leading to nuanced differences in estimated effects.

In terms of overall program impact, all three methods consistently estimated a

negative ATE, indicating that participation in the Head Start program was associated with lower average mathematics achievement as measured by the longitudinal IRT-based score (avg_MIRT). However, the magnitude of the estimated effects varied considerably. The PSM model, using 1:1 nearest neighbor matching with linear regression adjustment, yielded a conservative ATE of −1.2606. In contrast, the BART model produced a stronger estimate of −1.5421, and the Causal Forest model suggested an even more pronounced negative effect of −2.4340. These differences reflect the native adaptability of each technique: whereas PSM relies on linearity and balance with respect to a single scalar propensity score, BART and CF are able to handle high-order interactions and nonlinear functional forms, allowing them to model more sophisticated causal relationships contained in the data. A summary of these ATE estimates and corresponding CATE values across SES groups is presented in Table 1.

Table 1. ATE and CATE estimates with 95% intervals across three methods.

Method	ATE	95% Interval	CATE (Low SES)	CATE (Medium SES)	CATE (High SES)
PSM	−1.2606	[−1.6690, −0.8522]	—	—	—
BART	−1.5421	[−3.6752, 0.1863]	−0.9797	−1.4637	−2.1882
Causal Forest	−2.4340	[−2.4505, −2.4188]	−1.8117	−2.7929	−2.7018

Apart from approximating average effects, the capacity to identify treatment effect heterogeneity is a core distinction between procedures. BART and Causal Forest were the only ones to enable the estimation of Conditional Average Treatment Effects (CATEs) by Socioeconomic Status (SES), a crucial feature variable for education research. The BART model created a monotonically increasing gradient in which CATEs grew more negatively from low SES (−0.9797) to medium (−1.4637) and high SES (−2.1882) groups. This pattern suggests a diminishing marginal return to early intervention in progressively more advantaged children due to substitution effects or higher opportunity costs. The Causal Forest model, on the other hand, estimated the greatest impact in the medium SES group (−2.7929), followed very closely by the high (−2.7018) and low SES groups (−1.8117). This nonmonotonicity suggests that there may be underlying interactions, and these are perhaps more accurately captured by the data-adaptive splitting pattern of causal trees.

Lastly, model transparency and Individual Treatment Effects (ITEs) offered a further dimension of distinction between CF and BART. Both models offered unit-level ITEs, which were also plotted and summarized. The CF model's ITEs were found to be much more dispersed and left-skewed relative to BART and are actually sensitive to fine-grained heterogeneity. Furthermore, CF by itself yields interpretable feature importances, with WKSESL (socioeconomic status) as the most potent moderating variable of treatment impacts (accounting for 66.4% of model

importance), followed by public school attendance (S2KPUPRI) and food stamp receipt (P1FSTAMP). These findings underscore the overarching role that structural disadvantage plays in educational intervention effects.

In aggregate, the comparisons highlight that although all methods concur about the direction of the effect, they do not concur with flexibility in modeling, heterogeneity capture, and explanation depth. PSM offers a sharp and interpretable lower-bound standard, suitable for conservative estimation. Its ATE estimate is statistically significant, with a 95% confidence interval not containing zero, signifying a modest but stable negative effect. BART attains both model interpretability and performance by stable CATE patterns and ITE distributions that are robust, while its 95% posterior credible interval is slightly positive, reflecting greater uncertainty despite a more concentrated average effect. Causal Forest, by contrast, excels in uncovering rich heterogeneity and covariates modifying treatment response. Its bootstrap-based interval is narrow and fully negative, confirming the robustness of the estimated impact. These interval property variations must be seriously considered, along with assumptions in modeling, when choosing a causal inference technique depending on the character of research questions and the importance of heterogeneity in policy evaluation.

Collectively, these comparative findings reinforce the necessity of aligning estimation strategy with policy goals—whether aiming for interpretability, heterogeneity targeting, or statistical precision.

One reason for this surprising finding is the "substitution effect" or "fade-out" effect, widely discussed in the early intervention literature [19] [20]. More advantaged children often have access to other enrichment sources—private preschools, enriched home life, or high-quality elementary schools—that may make the marginal benefits of Head Start zero or even negative. In such cases, participation in a standardized public program may divert children away from more specialized or advanced alternatives, creating crowding-out effects [21]. This is consistent with earlier evidence of plateauing or lower achievement among more advantaged children following the intervention.

The heterogeneous treatment effects by SES also speak to the difficulty of early interventions. BART picked up on a monotonic gradient with progressively more negative effects by SES groups, while Causal Forest identified the most negative effect in the medium SES group. These patterns suggest different mechanisms at play—perhaps lower-SES children benefit minimally from Head Start due to a lack of alternatives, medium-SES children experience the most tension between program quality and aspirations, and high-SES children experience opportunity costs. These explanations must be investigated but suggest that Head Start is not benefiting strata equally.

Methodologically, the contrast between PSM, BART, and Causal Forest reveals the trade-off among interpretability, model flexibility, and sensitivity to heterogeneities. PSM was conservative but failed to detect large CATEs. BART produced stable and interpretable subgroup effects and smooth ITE distributions. CF was

most sensitive to covariate interactions and detected the substantive moderating effects of SES and school sector variables. Together, the methods provide a more comprehensive picture than could be achieved using any individual method alone.

Such findings have important policy implications. Rather than assuming homogeneous early intervention gains, program planners need to control for baseline conditions, access settings, and marginal benefit curves. An evidence-based targeting strategy whereby incremental assistance is allocated based on estimated individual treatment effects can increase equity and efficiency. Identification of subgroups underserved or overserved by the current design can further inform adaptive provision and reallocation of resources.

There are, nevertheless, several limitations to this study. First, the treated sample size was relatively small (N = 165), which limited the precision of subgroup analyses. Second, P1HSEVER as a proxy for treatment does not capture the quality or dose of Head Start exposure, and estimates could therefore be biased. Third, while we adjusted for key confounders, unobserved influences such as parental involvement or school environment could still influence both participation and outcomes. Finally, our analysis was restricted to academic achievement only; future research must incorporate socio-emotional and behavioral domains to facilitate a more holistic evaluation.

Looking ahead, future research would be assisted by combining ECLS-K with other datasets on school climate or parental behavior, examining dynamic treatment effects over time, and employing double-robust or targeted maximum likelihood estimators to strengthen causal identification. These extensions would further advance our understanding of how, for whom, and under what conditions early childhood interventions are most likely to succeed.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Holland, P.W. (1986) Statistics and Causal Inference. *Journal of the American Statistical Association*, **81**, 945-960. https://doi.org/10.1080/01621459.1986.10478354

[2] Rosenbaum, P.R. and Rubin, D.B. (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, **70**, 41-55. https://doi.org/10.1093/biomet/70.1.41

[3] Robins, J.M. and Rotnitzky, A. (1995) Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*, **90**, 122-129. https://doi.org/10.1080/01621459.1995.10476494

[4] Hill, J.L. (2011) Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, **20**, 217-240. https://doi.org/10.1198/jcgs.2010.08162

[5] Chipman, H.A., George, E.I. and McCulloch, R.E. (2010) BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, **4**, 266-298. https://doi.org/10.1214/09-aoas285

[6] Michelmore, K. and Dynarski, S. (2017) The Gap within the Gap: Using Longitudinal

Data to Understand Income Differences in Student Achievement. *AERA Open*, **3**, 1-18.

[7] Hahn, P.R., Murray, J.S. and Carvalho, C.M. (2020) Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, **15**, 965-1056. https://doi.org/10.1214/19-ba1195

[8] Sparapani, R.A., Logan, B.R., McCulloch, R.E. and Laud, P.W. (2016) Nonparametric Survival Analysis Using Bayesian Additive Regression Trees (BART). *Statistics in Medicine*, **35**, 2741-2753. https://doi.org/10.1002/sim.6893

[9] Kapelner, A. and Bleich, J. (2016) bartMachine: Machine Learning with Bayesian Additive Regression Trees. *Journal of Statistical Software*, **70**, 1-40. https://doi.org/10.18637/jss.v070.i04

[10] Linero, A.R. (2018) Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection. *Journal of the American Statistical Association*, **113**, 626-636. https://doi.org/10.1080/01621459.2016.1264957

[11] Taddy, M., Gardner, M., Chen, L. and Draper, D. (2016) A Nonparametric Bayesian Analysis of Heterogenous Treatment Effects in Digital Experimentation. *Journal of Business & Economic Statistics*, **34**, 661-672. https://doi.org/10.1080/07350015.2016.1172013

[12] Wager, S. and Athey, S. (2018) Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, **113**, 1228-1242. https://doi.org/10.1080/01621459.2017.1319839

[13] Denton, K. and West, J. (2002) Children's Reading and Mathematics Achievement in Kindergarten and First Grade. NCES 2002-125. U.S. Department of Education and National Center for Education Statistics.

[14] West, J., Denton, K. and Germino Hausken, E. (2000) America's Kindergartners: Findings from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99. NCES 2000-070. U.S. Department of Education and National Center for Education Statistics.

[15] Najarian, M., Sorongon, A.G., McManus, J.K., Walston, J.T. and Hagedorn, M.C. (2018) Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten-Fifth Grade Data File and Electronic Codebook (NCES 2018-032). U.S. Department of Education and National Center for Education Statistics.

[16] Tourangeau, K., Nord, C., Lê, T., Sorongon, A. and Najarian, M. (2015) Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten-Fourth Grade Data File and Electronic Codebook (NCES 2015-073). U.S. Department of Education and National Center for Education Statistics.

[17] Athey, S., Bayati, M., Doudchenko, N., Imbens, G. and Khosravi, K. (2020) EconML: A Python Package for ML-Based Heterogeneous Treatment Effect Estimation. Microsoft Research. https://github.com/microsoft/EconML

[18] Heckman, J.J. and Karapakula, G. (2019) Intergenerational and Intragenerational Externalities of the Perry Preschool Project. National Bureau of Economic Research. https://www.nber.org/papers/w25889

[19] Bailey, D., Duncan, G.J., Odgers, C.L. and Yu, W. (2017) Persistence and Fadeout in the Impacts of Child and Adolescent Interventions. *Journal of Research on Educational Effectiveness*, **10**, 7-39. https://doi.org/10.1080/19345747.2016.1232459

[20] Heckman, J.J. (2006) Skill Formation and the Economics of Investing in Disadvantaged Children. *Science*, **312**, 1900-1902. https://doi.org/10.1126/science.1128898

[21] Kline, P. and Walters, C. (2015) Evaluating Public Programs with Close Substitutes: The Case of Head Start. National Bureau of Economic Research. Working Paper Series. https://www.nber.org/papers/w21658