

Machine learning and propensity score matching for evaluating the effect of special education services on children's later math performances

Liyao Huang

Mathematics and Applied Mathematics, College of Sciences, Shanghai University,
Shanghai, China

claire0623_aboutmaths@shu.edu.cn

Abstract. The goal of the research is to analyze the effect of special education service brought to children's later math performance. The dataset is derived from Early Childhood Longitudinal Studies(ECLS) program and the children being studied come from diverse socioeconomic and racial/ethnic backgrounds. We applied PSM and modern machine learning methods including OLS, KNN, BART and MLP on the data in order to calculate the average treatment effect of special education services. It turned out that all the listed machine learning algorithms with comparatively low STD outperformed the traditional propensity score matching. KNN, BART and OLS excelled, offering much more stable calculations. The value of ATE computed through all the methods appeared below zero. By applying linear regression and PCA on all the influencing factors, the analysis revealed that the differences of some factors between the controlled and exposed groups led better math performances to appear usually in the absence of special treatment. Thus, special treatment effect led the trained model to predict lower scores, which finally caused the difference to be negative.

Keywords: Causal Inference, Machine Learning, Propensity Score Matching, Principal Component Analysis.

1. Introduction

To maintain the long goal of Early Childhood Longitudinal Studies(ECLS) program, the research is conducted primarily conducted to investigate and evaluate the effect of special education service brought to children's later math performance. The children being studied come from diverse socioeconomic and racial/ethnic backgrounds. In addition to the reception of special education services regarded as the exposure variable in our study, children varied in other factors associated with demographic, academic, school composition, family context health and parent rating of child, which can all possibly influence their math test scores and may also impact their likelihood of receiving the special education services.

Thus, the situation remains consistent with the theory of causal inference pointing out the difficulty of calculating the accurate treatment effect in non-randomized experiments. The most widely used method to minimize the impact of confounding variables is propensity score analysis. Nowadays, lots of machine learning algorithms are proposed to predict the outcome values if the controlled groups were to be treated or the exposed group had not received treatment. By predicting the outcomes, the problem

about randomization can be settled naturally. In our research, we applied Propensity Score Matching, Linear Regression, K-Nearest Neighbors, Bayesian Additive Regression Tree and Multi-Layers Perception to compute the average treatment effect(ATE). Besides, standard deviation(STD) of individual treatment effect is evaluated as well. By comparing the values of STD, more preferable methods stand out.

2. Literature Review

2.1. Traditional Causal Inference

Considering the potential causal relationship between observed features and past incidents, scientists proposed and developed the Rubin Causal Model to logically describe the causal inference, providing a meticulous and useful framework for experiments in causal inference. This model was applied in Holland and Rubin (1980) to analyze causal inference in retrospective, case-control studies used in medical research and in Holland and Rubin (1983) to analyze Lord's analysis of covariance" paradox[1]. In this model, we use the following notation: for each unit u in U , $Y(u)$ is the response variable(the outcome). "S" indicates the exposure of the unit during the experiment. For all the units, they would either receive certain treatment or not, which is usually regarded as controlled. The two conditions are denoted as follows: $S = t(\text{treatment})/S = c(\text{control})$. Moreover, $Y(u)$ are specified into $Y_t(u)$ and $Y_c(u)$, representing the response value of all units in U exposed to 2 conditions respectively. The effect of cause is defined as $Y_t(u) - Y_c(u)$. Since $Y_t(u)$ and $Y_c(u)$ cannot be derived simultaneously for the same unit, which is called the Fundamental Problem of Causal Inference. To quantify the effect of cause statistically and solve the problem, the Rubin Causal Model puts forward several assumptions and utilizes the expectation. Considering ATE(Average Treatment Effect) as the variable indicating the quantified average effect of the cause, then we have the following equation:

$$ATE = E(Y_t(u) - Y_c(u)) = E(Y_t(u)) - E(Y_c(u)) \quad (1)$$

By this equation, we can derive the exact value of the effect by calculating the average of response values for the treated condition and the controlled condition separately and doing the subtraction. As is mentioned above, the model is based on several kinds of assumption:

Temporal Stability: It is assumed that the causal effect of the treatment on the outcome is stable over time and across subgroups within the population. The value of $Y(u, c)$ does not depend on when the sequence "apply c to u then measure Y on u " occurs[1].

Causal Transience: The value of $Y(u, t)$ is not influenced by the previous exposure of u to c . [1]

Unit Homogeneity: Scientists in the laboratory prepare two units u_1, u_2 conscientiously so that they appear identical in all relevant aspects to be convinced that $Y(u_1, t) = Y(u_2, t)$ and $Y(u_1, c) = Y(u_2, c)$ always hold.

Independence: This assumption can be interpreted as two equations:

$$\begin{cases} E(Y_t(u)) = E(Y_t|S = t) \\ E(Y_c(u)) = E(Y_c|S = c) \end{cases} \quad (2)$$

Here, $E(Y_t(u))$ or $E(Y_c(u))$ represents the average response value of all the units when exposed to t or c , while $E(Y_t|S = t)$ or $E(Y_c|S = c)$ means the average response value of part of the whole U with the exposure to t or c . If (2) holds, we have

$$T = E(Y_t(u) - Y_c(u)) = E(Y_t(u)) - E(Y_c(u)) \quad (3)$$

$$T_{PF} = E(Y_t|S = t) - E(Y_c|S = c) \quad (4)$$

$$T = T_{PF} \quad (5)$$

By this formula, we can divide the original large sample into two parts ($S=t/S=c$) and observed the difference, recording the response value respectively.

Constant Effect: The effect of t on each unit u is regarded to be constant.

$$T = Y_t(u) - Y_c(u) = \text{constant}, \forall u \in U \quad (6)$$

If Unit Homogeneity holds, constant effect must be true, implying that Unit Homogeneity is a sufficient(stronger) condition for constant effect.

Randomized Controlled Trial&Exogenous Treatment Assignment: The units are randomly divided into two groups: one in the treatment group and one in the control group($S=t/S=c$). The treatment outcome is the outcome that would occur if the individual received the treatment, while the control outcome is the outcome that would occur if the individual had not received the treatment. The assignment of treatments must be exogenous to the potential outcomes—meaning that it must not be systematically related to the potential outcomes.

No Unmeasured Confounders: Unmeasured confounders are ensured to be excluded, meaning that all factors that could affect both the treatment assignment and the outcome have been measured and controlled for, since it is critical to avoid biased estimates of causal effects.

Based on the Rubin Causal Model and our ideal hypothesis, we compute $T = E(Y_t(u) - Y_c(u)) = E(Y_t(u)) - E(Y_c(u))$. That is because the sample and the treatment are independent and will not influence each other, usually guaranteed by randomized experiment.

However, in most cases, without appropriate randomization, some experiments involve samples that will correlate with the treatment systematically, resulting in the inequality between $E(Y_t(u))$ and $E(Y_t|S = t)$ and the equation of T will not hold. In such situation, we need to distinguish the disturbed one. So Propensity Score plays a key role by indicating the probability a unit receiving a particular treatment or intervention, given a set of observed characteristics or covariates.

$$\text{Propensity Score} \triangleq P(S_i = t|U_i) \quad (7)$$

It can be calculated through Logistic Regression in several computer language such as R[2]. Moreover, Bayesian Network can also applied through structure learning to preclude non-causal relations[3].

2.2. Machine Learning in Causal Inference

With the development of computer science and statistics, when dealing with tremendous amounts of data, machine learning provides efficient and precise methodology for researchers to analyze the experimental data. In Causal Inference, numerous ML techniques have been developed and applied in various studies. Traditional method such as ordinary least squares(OLS) for linear regression has been widely used in many studies, but they were limited when encountering more complicated datasets. More new developed ML methods such as K nearest neighbors(KNN), Bidirectional and Auto-Regressive Transformers(BART), Optimal Discriminant Analysis(ODA), Neural Network(NN),etc[4,5,6,7].One research utilized BART model to compute a causal estimand called TEA to estimate the number of adverse health events prevented by large-scale air quality regulations via changes in exposure to multiple pollutants, which outperformed standard parametric approaches[4]. Neural Networks were applied to a new causal model by building a ML proxy predictor of the conditional average treatment effect, followed by the consequence showing the practical use of the method and its advantage to avoid making strong assumptions[5]. Moreover, the application of both regression-based algorithm and optimal discriminant analysis to estimate multi-valued treatment effects using data from an intervention including three study groups revealed that ODA is a robust alternative to conventional regression-based models for estimating effects in multivalued treatment studies due to its insensitivity to skewed data and use of accuracy measures applicable to all prognostic analyses[6]. Generally speaking, different data structures correspond to different causal methods[7]. In most cases, including the researches listed above, new ML models surpass traditional regression model.

3. Methodology

3.1. Overview of Our Experiment

The flowchart of our whole work:

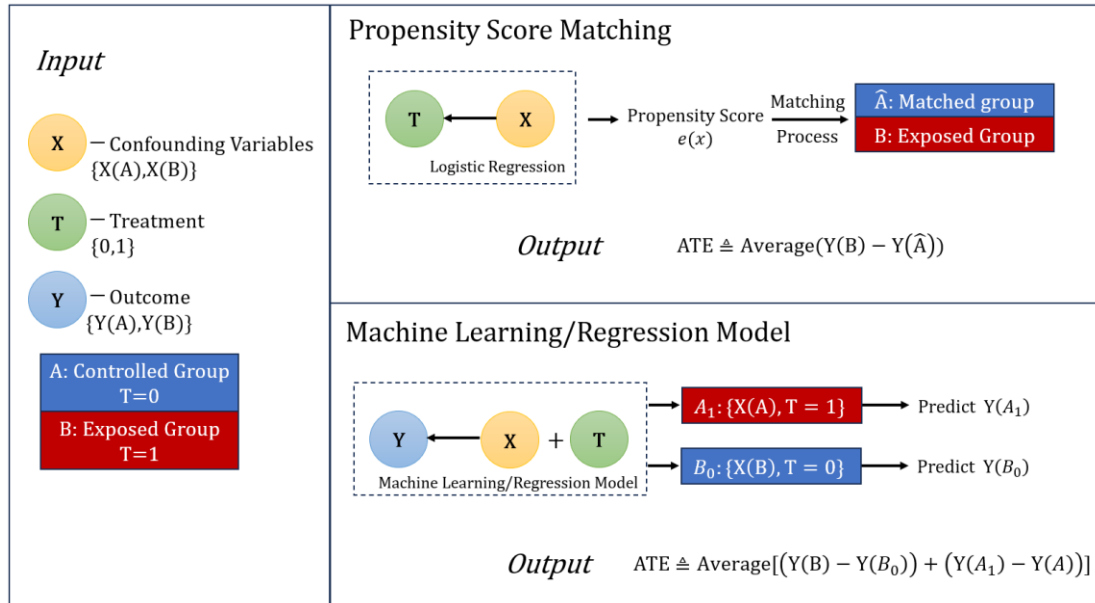


Figure 1. Flowchart of the Whole Research

3.2. Propensity Score Matching

As was mentioned before, in fact some experiments involve samples that will correlate with the treatment systematically, resulting in the inequality between $E(Y_t(u))$ and $E(Y_t|S=t)$. As a consequence, the difference between the outcomes of the controlled and the treatment groups does not merely shows the effect caused by the treatment but also those potentially contributed by the confounding variables.

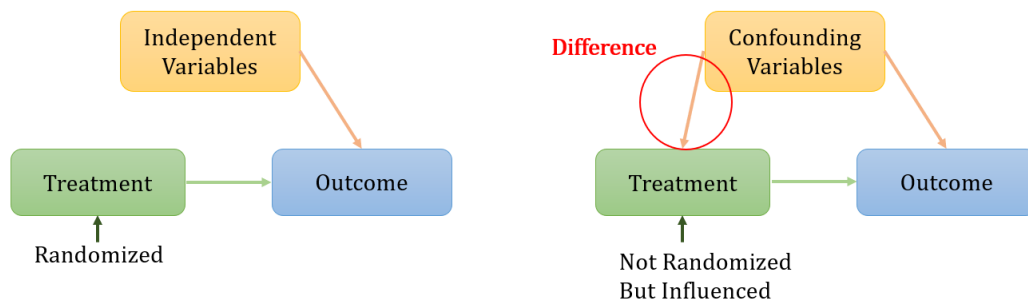


Figure 2. Relationships of Variables, Treatment and Outcome

Thus, Propensity Score Matching serves to construct an artificial control group by matching each treated unit with a non-treated unit of similar characteristics in order to reduce the error for estimating the treatment effect, which is a useful method in data analysis for estimating the impact of a program or event for which it is not ethically or logistically feasible to randomize. Let's review the definition of propensity score:

$$\text{Propensity Score} \triangleq P(S_i = t|U_i) \quad (8)$$

The score indicates the probability a unit receiving a particular treatment or intervention, given a set of observed characteristics or covariates. By knowing the definition, Propensity Score Analysis can be eventually divided into several steps. First, we need to collect data including all possible confounding variables that influence both the selection of treatment and the outcome variable. Then, having all the data, logistic regression model is commonly adopted as the propensity model to predict the propensity score equivalent to the probability of receiving the treatment given the confounders. By comparing the calculated propensity score, we manage to find a control record with the most similar propensity score for each example in the treatment group. Obtaining the match results, the quality of the matched records can be evaluated by comparing the similarity of the covariates. Finally, utilizing the new controlled group and the corresponding exposed group, the difference as well as the ATE can be computed.

3.3. Machine Learning Algorithms

3.3.1. Linear Regression: Ordinary Least Square

The most conventional method to analyze the correlations between the factors and the results is Linear Regression by Ordinary Least Square (OLS). The mentioned “Least Square” stands for the minimal square errors(SSE), which can be derived by adding the square differences of the observed and predicted values altogether.

Here we make the following assumptions:
the factor variables of all the trained objects:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

the observed values: $\mathbf{Y} = (y_1, y_2, \dots, y_m)^T$ the predicted value: $\hat{\mathbf{Y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_m)^T$

Since we assume the relation between the influential factors and the results can be expressed linearly, there exist $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_n)^T$ such that $\hat{\mathbf{Y}} = X\boldsymbol{\beta}$

By the definition of square errors, we have

$$SSE = \sum_{i=1}^m (\hat{y}_i - y_i)^2 = (\hat{\mathbf{Y}} - \mathbf{Y})^T (\hat{\mathbf{Y}} - \mathbf{Y}) = (X\boldsymbol{\beta} - \mathbf{Y})^T (X\boldsymbol{\beta} - \mathbf{Y}) \quad (9)$$

The aim is to find the optimal $\boldsymbol{\beta}^*$ to minimize SSE. Therefore, by the necessary optimal condition, it is required that $\nabla_{SSE}(\boldsymbol{\beta}^*) = 0$ i.e. $2X^T(X\boldsymbol{\beta}^* - \mathbf{Y}) = 0$

Finally, we can derive the optimal coefficients for the linear regression:

$$\boldsymbol{\beta}^* = (X^T X)^{-1} X^T \mathbf{Y} \quad (10)$$

More specifically, if only one factor is researched, we have $X = \begin{pmatrix} x_1 & x_2 & \cdots & x_m \\ 1 & 1 & \cdots & 1 \end{pmatrix}^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$

$$SSE = \sum_{i=1}^m (\hat{y}_i - y_i)^2 = \sum_{i=1}^m (\beta_1 x_i + \beta_0 - y_i)^2 \quad (11)$$

By the necessary optimal condition, we have

$$\begin{cases} \nabla_{SSE}(\beta_1^*) = 2 \sum_{i=1}^m x_i(\beta_1 x_i + \beta_0 - y_i) = 0 \\ \nabla_{SSE}(\beta_0^*) = 2 \sum_{i=1}^m (\beta_1 x_i + \beta_0 - y_i) = 0 \end{cases} \quad (12)$$

$$\begin{cases} \nabla_{SSE}(\beta_1^*) = 2 \sum_{i=1}^m x_i(\beta_1 x_i + \beta_0 - y_i) = 0 \\ \nabla_{SSE}(\beta_0^*) = 2 \sum_{i=1}^m (\beta_1 x_i + \beta_0 - y_i) = 0 \end{cases} \quad (13)$$

Solving the equations, the coefficients can be precisely calculated as:

$$\begin{cases} \beta_1^* = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^m (x_i - \bar{x})^2} \\ \beta_0^* = \bar{y} - \beta_1^* \bar{x} \end{cases} \quad (14)$$

$$\beta_0^* = \bar{y} - \beta_1^* \bar{x} \quad (15)$$

3.3.2. KNN Algorithm

Different from traditional methods, machine learning excels by its robust adaptability for various datasets. K-Nearest Neighbor Algorithm(KNN) is a widely used machine learning technique. After calculating the distance between input data and all the training sample points, the algorithm select K(the number of the selected points) nearest neighboring samples. For classification, the most common labels among the chosen neighbors is regarded as the predicted label of the input data. For regression, we assign the (weighted) average of the sample values of the neighbors as the predicted value of the input data. Obviously, the performance of the algorithm depends on parameter “K” and the distance metric[8].

Usually, the distance metric is determined among the following:

Euclidean distance($\|x\|_2$):

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum (y_i - x_i)^2} \quad (16)$$

Manhattan distance($\|x\|_1$):

$$d(\mathbf{x}, \mathbf{y}) = \sum |y_i - x_i| \quad (17)$$

Minkowski distance($\|x\|_p$):

$$d(\mathbf{x}, \mathbf{y}) = (\sum (y_i - x_i)^p)^{\frac{1}{p}} \quad (18)$$

3.3.3. Bayesian Additive Regression Tree

Combining the idea of regression and tree-based model, Bayesian Additive Regression Tree(BART) algorithm utilize sum of trees to approximate an unknown function f . Suggesting there is an unknown function f for predicting y based on the input x :

$$y = f(\mathbf{x}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \quad (19)$$

BART uses sums of trees $\sum g_j(\mathbf{x})$ where $g_j(\mathbf{x})$ denotes a regression tree to approximate the unknown function f :

$$f(\mathbf{x}) \approx h(\mathbf{x}) = \sum_{j=1}^m g_j(\mathbf{x}) \quad (20)$$

$$\Rightarrow y = h(\mathbf{x}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \quad (21)$$

Let T denote a binary tree and assume the tree has b terminal nodes, then $\mathbf{M} = (\mu_1, \mu_2, \dots, \mu_b)$ represents the parameter values corresponding to each of the terminal nodes. So, for a binary tree, each input value \mathbf{x} is associated with a single terminal node of T and is then assigned the μ_i value

corresponding to this node. Given the input \mathbf{x} and the tree T with a set of parameters \mathbf{M} , the regression tree can be denoted as $g(\mathbf{x}; T, \mathbf{M})$.

Based on this notation, the sum-of trees model can be mathematically written as:

$$y = \sum_{j=1}^m g(\mathbf{x}; T_j, \mathbf{M}_j) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \quad (22)$$

For each binary regression tree T_j , a set of parameters $\mathbf{M}_j = (\mu_{1j}, \mu_{2j}, \dots, \mu_{bj})$ is associated with the tree's terminal nodes.

The sum-of-trees model manage to incorporate both main effects and interaction effects. Large number of trees endows BART algorithm with robust predictive capabilities. For a fixed number of trees as m , $(T_1, \mathbf{M}_1), \dots, (T_m, \mathbf{M}_m), \sigma$ finally specify the sum-of-trees model, which is completed by imposing a prior over all the parameters aiming to limit the scale of individual tree preventing their effects from being exceedingly influential.

Firstly, it is assumed that the tree components (T_j, \mathbf{M}_j) are independent of each other as well as σ and the terminal node parameters of every tree are independent.

Then we have the following equations:

$$p((T_1, \mathbf{M}_1), \dots, (T_j, \mathbf{M}_j), \sigma) = \left[\prod_j p(T_j, \mathbf{M}_j) \right] p(\sigma) \quad (23)$$

$$= \left[\prod_j p(\mathbf{M}_j | T_j) p(T_j) \right] p(\sigma) \quad (24)$$

$$p(\mathbf{M}_j | T_j) = \prod_j p(\mu_{ij} | T_j) \quad (25)$$

According to equation (24) and (25), 1. the original specification problem is converted into the specification of $p(T_j), p(\mu_{ij} | T_j)$ and $p(\sigma)$. So the prior is separated into three parts: the T_j prior, the $\mu_{ij} | T_j$ prior, and the σ prior.

1) The T_j prior

The distribution on the splitting variable assignments and the distribution on the splitting rule assignment is the uniform prior on each binary tree distribution. In addition, the probability that a node at depth $d = (0, 1, 2, \dots)$ is nonterminal require further specification. The specification is given by: $\frac{\alpha}{(1+d)^\beta}$, $\alpha \in (0, 1)$ and $\beta \in [0, \infty)$. To keep individual tree components small, usually we adopt $\alpha = 0.95$ and $\beta = 2$.

2) The $\mu_{ij} | T_j$ prior

The output data y will be rescaled ranging from $y_{min} = -0.5$ to $y_{max} = 0.5$.

Then the prior is:

$$\mu_{ij} \sim \mathcal{N}(0, \sigma_\mu^2), \quad \sigma_\mu = \frac{0.5}{k\sqrt{m}} \quad (26)$$

The prior exerts the effort to shrink the parameter μ_{ij} toward zero. When k or m increases, the shrinkage to the μ_{ij} accelerates, making the prior tighter.

3) The σ prior

The prior exploit inverse chi-square distribution:

$$\sigma^2 \sim \frac{v\lambda}{\chi_v^2} \quad (27)$$

where v stands for the degree of freedom and λ is the scale.

We calibrate the prior for the degree of freedom v and scale λ for this purpose using a rough data-based overestimate $\hat{\sigma}$ of σ . Usually, there are two natural choices: naïve specification (the sample standard deviation of sample y) and the linear model specification (the residual standard deviation from a least squares linear regression of y on the original X). Then a value of v is picked between 3 and 10 to derive an appropriate shape and a value of λ is chosen such that the q th quantile of the prior on σ is located at $\hat{\sigma}$. The value of q is always considered such as 0.75, 0.90 or 0.99 to center the distribution below $\hat{\sigma}$.

For automatic use, Chipman et al. [9] recommend the default setting $(v, q) = (3, 0.90)$.

For the fixed tree number m , a fast and robust option is to set m as 200.

Given the observed data y of the sample value x , the Bayesian setup induces a posterior distribution:

$$p((T_1, \mathbf{M}_1), \dots, (T_j, \mathbf{M}_j), \sigma | y) \quad (28)$$

Through approaches such as a backfitting MCMC algorithm proposed by Chipman et al. [9] and Particle Gibbs proposed by Lakshminarayanan et al. [10], the parameters can be specified.

As a consequence, the BART model can successfully output a posterior mean estimate of $f(\mathbf{x}) = E(y | \mathbf{x})$.

Table 1. Iteration of BART Algorithm

The iteration of the algorithm
Initialize the model:
$(T_1^{(1)})^{\mathbf{M}_1^{(1)}}(\mathbf{x}) = (T_2^{(1)})^{\mathbf{M}_2^{(1)}}(\mathbf{x}) = \dots = (T_m^{(1)})^{\mathbf{M}_m^{(1)}}(\mathbf{x}) = \frac{1}{nm} \sum_{i=1}^n y_i$
Compute $\hat{f}^{(1)}(\mathbf{x}) = \sum_{t=1}^m (T_t^{(1)})^{\mathbf{M}_t^{(1)}}(\mathbf{x})$
Begin iteration :
For $k = 2, 3, \dots, K$
For $j = 1, 2, 3, \dots, m$
Compute $R_{-j}^{(k)} = y - \sum_{t < j} (T_t^{(k)})^{\mathbf{M}_t^{(k)}}(\mathbf{x}) - \sum_{t > j} (T_t^{(k-1)})^{\mathbf{M}_t^{(k-1)}}(\mathbf{x})$
Sample $p(T_j R_{-j}^{(k)}, (\sigma^{(k-1)})^2)$, derive $T_j^{(k)}$
Sample $p(\mathbf{M}_j T_j, R_{-j}^{(k)}, (\sigma^{(k-1)})^2)$, derive $\mathbf{M}_j^{(k)}$
Compute $\hat{f}^{(k)}(\mathbf{x}) = \sum_{t=1}^m (T_t^{(k)})^{\mathbf{M}_t^{(k)}}(\mathbf{x})$
Compute $\varepsilon^{(k)} = y - \hat{f}^{(k)}(\mathbf{x})$
Sample $p(\sigma^2 T_1^{(k)}, \mathbf{M}_1^{(k)}, \dots, T_m^{(k)}, \mathbf{M}_m^{(k)}, \varepsilon^{(k)})$, derive $(\sigma^{(k)})^2$
Final BART model:
$\hat{f}(\mathbf{x}) = \frac{1}{K-L} \sum_{k=L+1}^K \hat{f}^{(k)}(\mathbf{x})$

3.3.4. Neural Network: Multi-Layer Perception

Among machine learning techniques, Neural Network is a popular method. Its archetypical model Multi-layer Perceptron is a supervised learning algorithm that learns a function $f: R^m \rightarrow R^o$ by training on a dataset where m is the number of dimensions for input and o is the number of dimensions for output.

Given a set of features $\mathbf{X} = x_1, x_2, \dots, x_m$ and a target y , it can learn a non-linear function approximator for either classification or regression. The model consists of multiple non-linear layers usually called the hidden layers.

Without loss of generality, we can assume now we have input data $\mathbf{X} = x_1, x_2, \dots, x_m$ and the hidden layer $\mathbf{H} = h_1, h_2, \dots, h_o$. The model serves to compute the target value y .

For each element x_i , it has a set of weight values $\{w_{i1}, w_{i2}, \dots, w_{io}\}$ corresponding to the calculation of the elements in the hidden layer:

$$h_j = f \left[\left(\sum_{i=1}^m w_{ij} x_i \right) + b \right], j = 1, 2, \dots, o \quad (29)$$

where f is a function which can be either linear or non-linear (more often).

Three types of commonly adopted functions:

1) relu:

$$y = \max(0, x) \quad (30)$$

2) tanh:

$$y = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (31)$$

3) sigmoid:

$$y = \frac{1}{1 + e^{-x}} \quad (32)$$

Then for the hidden layer \mathbf{H} , each element has a set of weight values for the next layer.

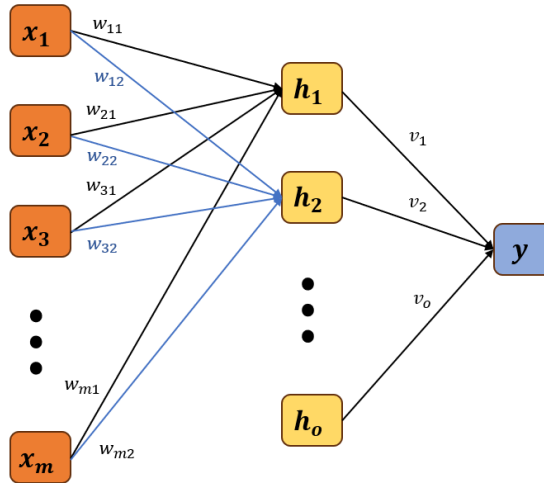


Figure 3. Working Mechanism of MLP

If the hidden layer is the last layer before the output data as the graph shows, then we can calculate the target value y :

$$y = \left(\sum_{i=1}^o v_i h_i \right) + s \quad (33)$$

Since there are several numerous parameters involved in the computing process, they need to be specified. They are specified to obtain the minimized error for the training datasets.

If we denote the predicted target values for the training data as $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$ where n is the sample size and consider $\mathbf{y} = (y_1, y_2, \dots, y_n)$ as the observed data.

Then the error can be described as:

$$\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (34)$$

Since the predicted value $\hat{\mathbf{y}}$ was determined by parameter $(w_{ij})_{m \times o}, b, v_i$ and s , we can also denote the error as a loss function: $L(\phi(\mathbf{x}; w, b, \mathbf{v}, s), y)$. Then to derive the most satisfying parameters, we need to solve the following optimization problem:

$$\min_{w, b, \mathbf{v}, s} L(\phi(\mathbf{x}; w, b, \mathbf{v}, s), y) \quad (35)$$

Usually for optimization, there are approaches for us to adopt to compute the solutions, such as Stochastic Gradient Descent(SGD), momentum method, Nesterov Accelerated Method and AdaGrad/RMSProp/AdaDelta/Adam Algorithms.

4. Experiment and Result

4.1. Data Description

In our experiment, the data originally come from the Early Childhood Longitudinal Studies (ECLS) program, Kindergarten Class of 1998-1999 (ECLS-K). The dataset contains a wide range of family, school, community, and individual factors related to children's development, early learning, and performance in school. To be more detailed about our evaluation about the causal inference, the covariates consist of factors related to demographic, academic, school composition, family context, health and parent rating of child. The treatment here is indicated by the exposure variable "Special Education Services" which is denoted as F5SPECS. For the purpose of the ECLS-K, children with disabilities are those who meet the federal eligibility requirements for participation in special education programs or services. All children with disabilities are expected to have an Individualized Education Plan (IEP), an Individualized Family Service Plan (IFSP), or a 504 Plan on file with the school district as it is a required component of the eligibility process.

After data pre-processing, the whole datasets consisted of 7362 cases, 429 of which received special education services. Whether a child was regarded as a recipient of special education services depended on their special education status gathered from school administrative records from the spring of 2002[2]. According to the past analysis in an issue brief by IES describing Demographic and School Characteristics of Students Receiving Special Education in the Elementary Grades[11], for the cohort of students beginning kindergarten in 1998, specific learning disabilities and speech or language impairments were the most prevalent primary disabilities over the grades studied. Higher percentages of boys than girls and of poor students than non-poor students received special education. In each grade studied, public schools reported higher percentages of students receiving special education than did private schools.

Meanwhile, the ECLS-K Revised IRT scaled math achievement test score ranging from 50.9 to 170.7 is considered the outcome variable.

4.2. Result and Experiment of Causal Inference

In the coding process, we mainly chose python to run the models through different models to calculate the average treatment effect and the standard deviation of individual treatment effect. By comparing the the standard deviation of individual treatment effect, we can assess the performance of all the applied methods. The computing process can be categorized into two parts composed of different steps to solve the fundamental problem of calculate treatment effect:

1) In the method of Propensity Score Matching, we first adopt logistic regression based on the covariates and the exposure variable to compute the probability of each individual receiving the treatment with their covariates information. By matching the exposed individual($n=429$) with the controlled one having the closest score to the exposed one and then subtracting the controlled score from the exposed score, we can derive individual treatment effect and compute ATE and the standard deviation of ITE($n=429$).

2) In the machine learning part, by applying the packages that are already created for logistics regression, OLS, KNN, BART, MLP on the whole datasets, we trained models predicting the outcomes based on the covariates and the exposure variables. Then we reversed the exposure variables(changing 0 to 1, changing 1 to 0) and evaluate new outcomes. Eventually, we obtain the response variables of the whole datasets($n=7362$) under both circumstances. Subtracting the controlled score from the exposed score, we derive the individual treatment effect and compute ATE and the standard deviation(STD) of ITE($n=7362$).

Table 2. ATE & Standard deviation of Individual Treatment Effect computed by Listed Algorithms

Methodology	MLP (linear-tanh-linear)	KNN	OLS	BART	PSM
ATE	-22.35	-0.22	-7.08	-6.12	-20.65
STD	23.90	14.23	15.64	15.16	34.66

Surprisingly, the average treatment effects of Special Education Services calculated turn out to be all minus. Besides, the standard deviation revealed that KNN, OLS and BART are more stable with better performances in calculating ATE. Whereas, MLP and PSM brought higher STD values indicating the fluctuation of the individual treatment effect among the datasets.

4.3. Factor Analysis

It may seem surprising that the treatment effect evaluated turned out to be minus. Since the datasets containing 7363 sets of data only have 429 test takers experiencing special education services. So the minority of the exposed individuals suggesting some hidden selection bias serves as a possible reason explaining why the average treatment effects calculated by all methods are minus.

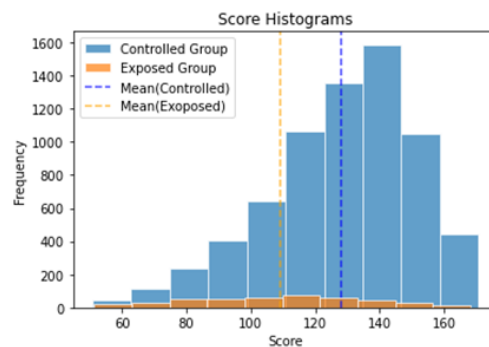


Figure 4. Histograms of Math Scores of Controlled and Exposed Groups

Table 3. Mean and Standard Deviation of Math Scores of Controlled and Exposed Groups

	Mean	STD
Controlled Group	128.19	22.67
Exposed Group	108.97	26.77

According to the histograms of the controlled and the exposed group, we can easily conclude that scores of exposed group are symmetrically distributed. Comparatively, the controlled scores are skewed

toward higher scores. Consequently, the mean score of controlled group is originally higher than the one of the exposed score. Thus, we can probably speculate that the exposed group receiving special educational service have potential disadvantage inherently such as intelligence disabilities or psychological problems in terms of math performance. Meanwhile, when the machine learning algorithms begins to learn the datasets, the higher scores are often appeared without the special treatment. Vice Versa, the special treatment is always accompanied by lower math scores, leading the trained model to predict higher/lower scores when treatment variable turns into 0/1, which finally caused the difference to be minus. What factors actually cause the outcomes to be different? Here we adopt the traditional regression method OLS to compute the results showing the influencing coefficient of all the factors. Compare with other machine learning algorithms, OLS excels in its ability to quantify the relationships of the factors and outcomes.

Table 4. Contributing Parameters of Variables to Math Scores(OLS)

<i>OLS</i>	GENDER	WKWHITE	WKSESL	RIRT	MIRT	S2KPUPRI	P1EXPECT
Coefficient	6.11	2.33	1.93	-0.03	1.19	5.50	0.43
<i>OLS</i>	P1FIRKDG	P1AGEENT	apprchT1	P1HSEVER	chg14	avg_RIRT	avg_MIRT
Coefficient	12.02	-0.72	2.97	-2.97	0.54	0.11	-0.14
<i>OLS</i>	avg_SES	avg_apprchT1	S2KMINOR	P1FSTAMP	ONEPARENT	STEPPARENT	P1NUMSIB
Coefficient	3.04	-1.65	-0.57	-1.55	-1.32	-0.12	-0.09
<i>OLS</i>	P1HMAFB	WKCAREPK	P1EARLY	wt_ounces	C1FMOTOR	C1GMOTOR	P1HSCALE
Coefficient	0.21	-1.11	0.04	0.04	1.69	-0.19	0.22
<i>OLS</i>	P1SADLON	P1IMPULS	P1ATTENI	P1SOLVE	P1PRONOU	P1DISABL	F5SPECS
Coefficient	0.23	-0.15	-0.55	-0.86	-0.67	0.81	-7.08

The table illustrates the highly influencing factors with high absolute values of the coefficients include whether the child go to public school, the child's genders, whether the child receive special education services and whether the child is a first-time kindergartener.

Except the factor of special education services that is negatively related to the math scores, the other three factors are all positively related to the output variables and the identity of first-time kindergartner contributes the most.

Having found the outstanding influential factors and their relations with the final outcomes, let's check those factors 'distribution in the controlled and exposed group respectively.

Table 5(a). Proportions of Public School Students in Controlled and Exposed Groups

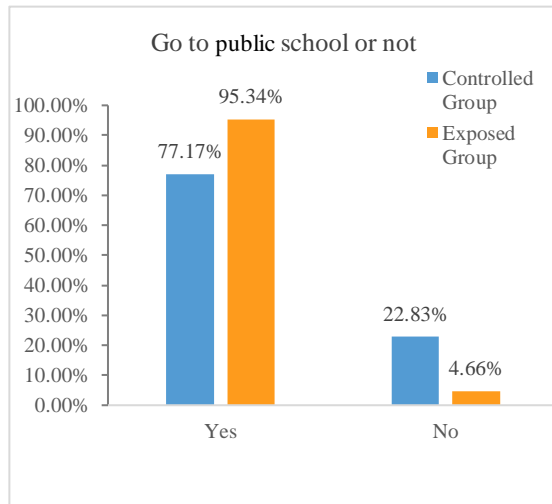


Table 5(b). Proportions of Male and Female Students in Controlled and Exposed Groups

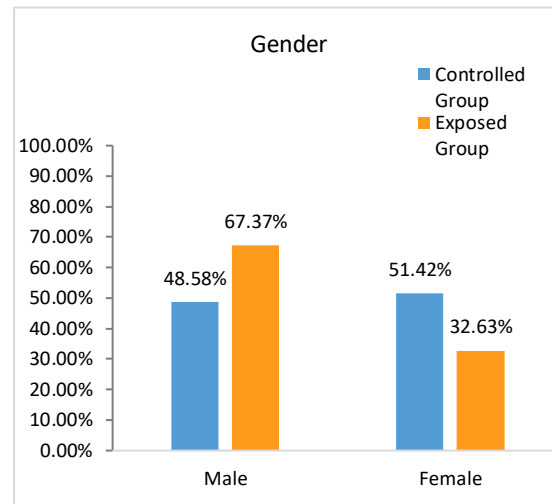
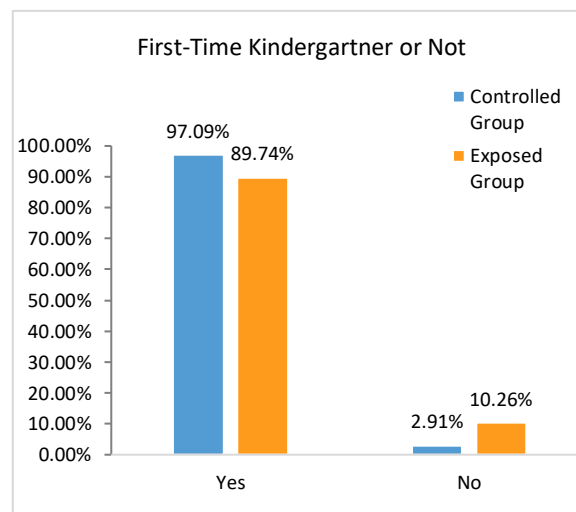


Table 5(c). Proportions of First-Time Kindergartner Students in Controlled and Exposed Groups



Since all the influential factors are recorded by binary response, here we calculated the frequency of two categories of the answer{0,1} in the controlled and exposed group in order to distinguish the differences.

For the factor of special education services causing decline of the score, apparently the exposed group are 100% receiving the services. The most influential factor which is the identity of first-time kindergartner has higher frequency among the controlled group than the exposed group, which also help explain why the controlled group performed better in the math test. Whereas, the other positively-related factors consisting of gender and public school turned out to be in favor of the exposed groups' scores. Nevertheless, with the largest influential parameter of first-time kindergartner and isolation from treatment services, the controlled group still outperformed the exposed one.

To dig and explore the more distinguished characteristics of the factors, Principle Components Analysis(PCA) is also adopted here on the whole datasets outputting continuous component variables. Applying Principle Components Analysis(PCA) on all the characteristics of individual including the

covariates and the exposure variable, we found out that 6 principle components accounted for 95% of all the components.

Table 6. Constituting Parameters of Variables in PC1~PC6(PCA)

Variables	PC1	PC2	PC3	PC4	PC5	PC6
GENDER	0.00	0.00	0.00	-0.06	-0.08	0.00
WKWHITE	0.00	0.00	-0.02	-0.71	-0.59	0.01
WKSESL	0.01	0.01	0.00	-0.06	0.00	0.00
RIRT	0.00	-0.01	-0.04	0.47	-0.02	0.02
MIRT	0.01	0.01	-0.01	-0.50	0.75	0.00
S2KPUPRI	0.00	0.01	0.04	0.00	0.05	0.00
P1EXPECT	0.00	0.01	0.00	0.09	-0.23	0.00
P1FIRKDG	-0.01	0.05	0.02	0.12	-0.13	-0.01
P1AGEENT	0.05	0.02	0.00	-0.01	0.08	0.01
apprchT1	-0.01	-0.01	-0.01	0.00	0.03	0.00
P1HSEVER	-0.02	0.16	0.03	-0.01	0.01	0.02
chg14	0.01	0.00	-0.09	-0.01	0.01	0.01
avg_RIRT	-0.04	0.02	0.21	-0.01	-0.01	-0.02
avg_MIRT	0.14	-0.03	-0.02	0.01	0.01	0.00
avg_SES	-0.14	-0.01	-0.03	-0.01	-0.01	0.02
avg_apprchT1	-0.05	-0.02	-0.37	0.00	0.00	0.04
S2KMINOR	0.10	0.04	0.75	0.00	-0.01	-0.11
P1FSTAMP	0.19	-0.01	-0.27	0.00	0.01	-0.05
ONEPARENT	0.31	-0.05	0.04	0.00	-0.01	0.00
STEPPARENT	-0.55	-0.06	0.06	0.00	0.02	-0.10
P1NUMSIB	0.68	0.00	-0.03	0.01	-0.01	0.02
P1HMAFB	-0.17	0.26	-0.10	0.00	0.00	0.03
WKCAREPK	0.05	-0.06	-0.19	0.00	0.00	-0.94
P1EARLY	-0.05	0.28	-0.07	0.00	0.00	-0.03
wt_ounces	0.05	0.04	-0.11	0.00	0.00	0.05
C1FMOTOR	-0.03	-0.73	0.03	0.00	0.00	0.07
C1GMOTOR	-0.04	-0.41	0.03	0.00	0.00	-0.04
P1HSCALE	0.02	-0.01	-0.01	0.00	0.00	-0.03
P1SADLON	0.02	0.31	0.14	0.00	0.00	-0.09
P1IMPULS	0.02	-0.03	0.26	0.00	0.00	-0.23
P1ATTENI	0.00	-0.11	0.04	0.00	0.00	0.00
P1SOLVE	0.05	0.00	-0.04	0.00	-0.01	0.10
P1PRONOU	0.00	-0.06	0.01	0.00	0.00	-0.01
P1DISABL	0.02	0.03	-0.02	0.00	0.00	0.08
F5SPECS	0.00	-0.02	0.02	0.00	0.00	0.01

The parameters of the principal components in the table shows that among all the features of the individuals, the group of factors about family context is embedded in the 6 components to the greatest extent followed by health, school composition and academic progressively.

All those components mainly indicates the variances of individuals.

Table 7. Contributing Parameters of PC1~PC6 to Math Scores(OLS)

OLS	PC1	PC2	PC3	PC4	PC5	PC6
Coefficient	-2.53	-0.15	-0.15	-0.12	1.89	3.25

Just like the way we had selected primary influencing factors before PCA, we utilized OLS for regression analysis and chose PC1,PC2 and PC6 with high absolute values of coefficients as our main investigated variables. The coefficient table reveals that PC5 and PC6 play significant role in enhancing the math scores while PC1 affects negatively.

Table 8(a). Histograms of PC1 of Controlled and Exposed Groups

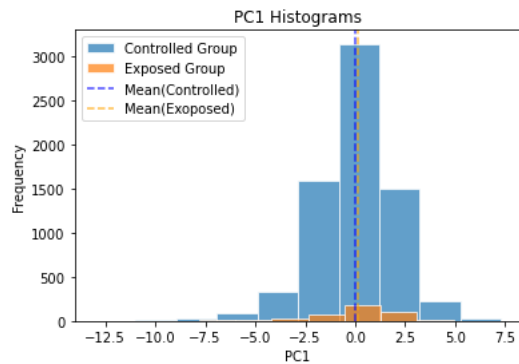


Table 8(b). Histograms of PC5 of Controlled and Exposed Groups

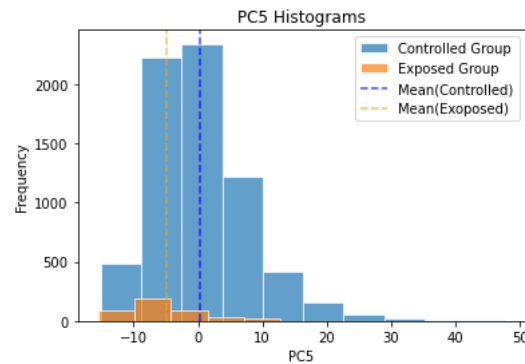
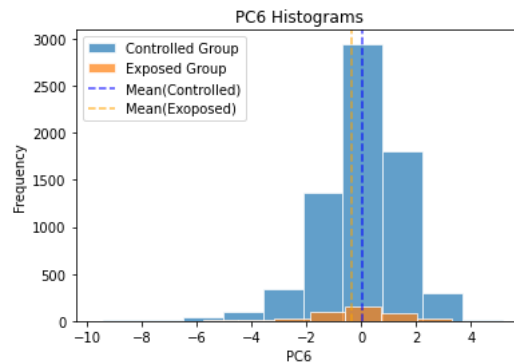


Table 8(c). Histograms of PC6 of Controlled and Exposed Groups



By comparing the distribution and the mean value of each group, we found that the controlled group has relatively higher values of PC5 and PC6 and lower value of PC1. The variations of all these three components contribute to greater outputs of the controlled group. The results derived from the projected data after PCA show more consistent and less antagonistic impact of the factors, which help verify the effectiveness of PCA adoption on our datasets.

5. Conclusion

First, in terms of the standard deviation of the ITE, all the machine learning algorithms with comparatively low STD outperformed the traditional propensity score matching. Among those machine learning algorithms, KNN, BART and OLS excelled, offering much more stable calculations.

Moreover, the values of ATE computed through all the methods turned out to be below zero. The results of factor analysis revealed that kids' math scores has intensively positive relationships with the identity of first-time kindergartners and Kindergarten Math Scores while strongly impaired by the

Number of Siblings and Non-parental Pre-K Child Care. The controlled group boasted more first-time kindergartner and higher Kindergarten Math Scores with smaller Number of Siblings and less Non-parental Pre-K Child Care, reasonably causing the greater performance in the math test scores. The machine learning algorithms learned that higher scores often appeared without the special treatment. Thus, special treatment effect led the trained model to predict lower scores, which finally caused the difference to be negative.

References

- [1] Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396), 945–960. <https://doi.org/10.1080/01621459.1986.10478354>
- [2] Keller, B., & Tipton, E. (2016). Propensity score analysis in R: A software review. *Journal of Educational and Behavioral Statistics*, 41(3), 326-348. <https://doi.org/10.3102/1076998616631744>
- [3] Sieswerda, M., Xie, S., et al. (2023). Identifying confounders using bayesian networks and estimating treatment effect in prostate cancer with observational data. *JCO Clinical Cancer Informatics*, 7, e2200080. <https://doi.org/10.1200/CCI.22.00080>
- [4] Nethery, R. C., Mealli, F., Sacks, J. D., & Dominici, F. (2019). Causal inference and machine learning approaches for evaluation of the health impacts of large-scale air quality regulations. *arXiv preprint arXiv:1909.09611*. <https://doi.org/10.48550/arXiv.1909.09611>
- [5] Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2018). *Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India* (No. w24678). National Bureau of Economic Research. <https://www.nber.org/papers/w24678>
- [6] Linden, A., & Yarnold, P. R. (2016). Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *Journal of Evaluation in Clinical Practice*, 22(6), 875-885. <https://doi.org/10.1111/jep.12610>
- [7] Parikh, H., Varjao, C., Xu, L., & Tchetgen, E. T. (2022, June). Validating causal inference methods. In *International conference on machine learning* (pp. 17346-17358). PMLR.
- [8] *What is the K-nearest neighbors algorithm?*. IBM. (2021, October 4). <https://www.ibm.com/topics/knn>
- [9] Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. <https://doi.org/10.1214/09-AOAS285>
- [10] Lakshminarayanan, B., Roy, D., & Teh, Y. W. (2015, February). Particle Gibbs for Bayesian additive regression trees. In *Artificial Intelligence and Statistics* (pp. 553-561). PMLR.
- [11] U.S. Department of Education. (2007a, July). *Demographic and School Characteristics of Students Receiving Special Education in the Elementary Grades*. AIR. <https://www.air.org/project/early-childhood-longitudinal-studies-ecls>