



Intergenerational education mobility: A machine learning perspective

 Xiang Gao¹
 Dongdong Li²
 Weige Huang^{3*}

¹Research Center of Finance, Shanghai Business School, 2271 West Zhong Shan Rd., Shanghai 200235, China.

Email: gaoxiang@sbs.edu.cn

²Office of Academic Research, Guangxi University of Foreign Languages, 19 Wuhu Avenue, Nanning 530222, China.

Email: gzky2022@163.com

³Wenlan School of Business, Zhongnan University of Economics and Law, 182 Nanhu Ave., Wuhan 430073, China.

Email: weige_huang@zuel.edu.cn



(+ Corresponding author)

ABSTRACT

Article History

Received: 6 November 2022

Revised: 30 December 2022

Accepted: 12 January 2023

Published: 27 January 2023

Keywords

Intergenerational education
Machine learning mobility
Random forest
SHAP
XGBoost.

JEL Classification

C40; D31; I20; J62.

By exploiting XGBoost and SHAP algorithms, this paper aims to reveal the importance of understanding the nexus among parental characteristics for intergenerational mobility in educational attainment, especially from the statistical learning perspective. Consistent with previous findings, both parents' income and head's education are positively correlated with child's income. We also show strong intergenerational education mobility for low- and high-income families. However, there exists a negative relationship between head and child's education for the middle-income families. Unlike conventional wisdom, we find that the income of highly educated parents tends to negatively associate with child's education and the opposite happens with poorly educated parents. Moreover, for white and black children, their parents' income will adversely affect child's education, but this effect turns out to be positive for children of other races. Our paper hence suggests the consideration of ethnicity and family wealth in conjunction when making education effectiveness facilitation policies.

Contribution/Originality: This research is among the first to apply statistical learning algorithms and methods to investigate education research field topics. The finding complements the extant literature by documenting that the income of highly educated parents might exhibit a negative correlation with their child's education.

1. INTRODUCTION

There has been abundant research in intergenerational education mobility across time, geography, individual or family characteristics, and macroeconomic conditions. Fletcher and Han (2018) document differences in education mobility during 1982-2004 among U.S. states. They find mobility fluctuation over time and identify a lack of increase in mobility in Southern U.S. Leone (2019) take one step further to estimate the worldwide variation for a longer period. The conclusion is that the mobility gap between the rich and poor countries has increased over time and the intergenerational persistence in education is strong in least-developed countries. Other studies also study this issue in particular countries (Aydemir & Yazici, 2019; Azomahou & Yitbarek, 2021) by focusing on race and gender (Ferrare, 2016) as well other factors (Engzell & Troupf, 2019; Jungert, Levine, & Koestner, 2020; Turcotte, 2011) and market conditions such as the financial development (Russino, 2018). But to the best of our knowledge, no analysis has been conducted to determine the interactive effects of different family features simultaneously. With the advances in machine learning, we can apply existing methodologies to uncover hidden relationships between them and explore the importance of features and relevant interaction effects.

This paper uses the eXtreme Gradient Boosting (XGBoost) and Shapley Additive exPlanations (SHAP) to study the importance of family characteristics in predicting a child's educational attainment and the interactions between these characteristics. XGBoost is a machine learning method used to model the relation between a child's education (our target variable) and family characteristics. It is essentially an effective way to put gradient-boosted decision trees to practical use. Similar to a tree structure, a decision tree also has a trunk (internal nodes), leaf nodes (end nodes), and a root node (topmost node). Decision tree algorithms typically employ straightforward rules to begin at the root node and branch out through internal nodes before reaching the leaves. Gradient-boosted decision trees, on the other hand, follow an ensemble learning method by employing a series of decision trees, with each decision tree influencing the next to enhance the model and create a robust learner (Chen & Guestrin, 2016; Parsa, Movahedi, Taghipour, Derrible, & Mohammadian, 2020).

Having said that, XGBoost's results can be interpreted using SHAP. Generally speaking, the reason why a model generates a particular prediction can be just as important as its accuracy. However, complex models like ensemble or deep learning models, which even experts have trouble understanding, frequently achieve the highest accuracy for huge datasets in the big data era, putting accuracy and interpretability at odds. To mitigate such concerns, a number of approaches have been proposed to assist users in interpreting the predictions made by complicated modeling setups. However, it remains unsolved how these approaches relate to one another or when we should prefer one approach over another. Fortunately, Lundberg & Lee (2017) propose SHAP, a unified framework for interpreting theoretical predictions, as a potential solution to the above "how" and "when" questions.

As for main results, we find that parents' income is a very important family characteristic for predicting a child's education but the most important variable for predicting child's education is head's education instead of parents' income. Consistent with extant literature, both parents' income and education are positively correlated with child's income. We also show that for low- and high-income families, there exists strong intergenerational education mobility. However, for the middle-income family parents' and child's education is negatively correlated. Unlike conventional wisdom, we find that the income of highly educated parents tends to negatively associate with child's education and the opposite happens with poorly educated parents. Moreover, for white and black children, their parents' income will adversely affect child's education, but this effect turns out to be positive for children of other races.

Our paper differs from recent related researches in the following aspects. The previous literature either explicitly or implicitly assumes that there exists a non-linear relation between parents' and child's education and the focus lies on studying the mean or other distributional effects of head's education on child's education. The evidence of the existence of distributional effects implies non-linear relations between a child's education and head's education. Therefore, a more complicated model is necessary in order to capture the non-linearity. In this paper, we use a machine learning method, the XGBoost model, which has been shown to work well (Lundberg et al., 2020). This is the first difference between this paper and the researches cited above. The second difference is that we also focus on the variable importance analysis (i.e., exploring which variables are more important for predicting target variable), which is what the extant studies have ignored. The last and very important difference between this research and other intergenerational education mobility studies is that this paper examines the interaction effects between other characteristics and head's education and finds very interesting results.

2. METHODOLOGY

This section provides an overview of XG Boost, which is used to model intergenerational education mobility, and SHAP, which is used to explain the model output. To begin with, we introduce the regression decision tree in machine learning for predicting non-binary numeric outcomes like the number of years a child spent at school in our setup. A decision tree has a flow-chart-like tree structure with one root node (corresponding to the dependent variable $y \in \mathbb{R}$), several layers of internal nodes, and T leaf nodes (corresponding to a vector of independent variables $\mathbf{x} = \{x^1, \dots, x^m\} \in \mathbb{R}^m$ or features in computer science language). Since there exist many alternative tree

structures, let $\mathcal{F} = \{f(\mathbf{x}) = \mathbf{w}_{q(x)}\}(q: \mathbb{R}^m \rightarrow T, \mathbf{w} \in \mathbb{R}^T)$ represent the space of trees, where each $f(\mathbf{x})$ is an independent structure q that maps any observation to its corresponding leaf weight \mathbf{w} .

Any single tree structure can be flawed with over-fitting, bias, and variance errors. We hence enhance such algorithm by either averaging out the end-note solutions of a large number of independent trees (random forest) or by building one tree on another additively along the way (gradient boosting originated from Friedman (2001)). This paper chooses the latter improvement to predict child education with family features because gradient boosting is proved to perform better if the data is unbalanced and has less noise. And XGBoost is the most efficient and accurate implementation of our selected gradient boosting algorithm.

Mathematically, for the i -th observation in the dataset, when our goal is to predict \hat{y}_i using \mathbf{x}_i and K additive functions,

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^K f_k(\mathbf{x}_i), f_k \in \mathcal{F},$$

The following Equation 1 is minimized:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \tag{1}$$

Where l is a loss function that tells how much different \hat{y}_i and y_i are from one another; Additionally, a penalty function that reduces model complexity is $\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\mathbf{w}\|^2$. The machine with regularized objectives naturally selects the simplest and most accurate model.

Tackling the above minimization issue gives us ideal weight w_j^* of leaf j as well as the comparing ideal incentive for $\tilde{\mathcal{L}}^{(t)}(q)$. Although calculating these values is a difficult task, the following formula is typically used to evaluate split candidates:

$$\mathcal{L}_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma. \tag{2}$$

In Equation 2, I_L and I_R are the instance sets of left and right nodes following the split, and $I = I_L \cup I_R$. $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial_{\hat{y}^{(t-1)}}^2 l(y_i, \hat{y}^{(t-1)})$ are the first- and second-order gradient statistics on the loss function, respectively. We use SHAP to interpret this model's output. See also Lundberg, Erion, and Lee (2018) and Lundberg et al. (2020) to learn more. As proposed by Lundberg and Lee (2017), SHAP estimates the contribution of each feature using game theory (Štrumbelj and Kononenko (2014)) and local explanations (Ribeiro, Singh, & Guestrin, 2016).

The target for prediction based on observation \mathbf{x}_i is given by the model of $v(\mathbf{x})$. A marginal contribution is the SHAP contribution ψ_m of each feature m on $v(\mathbf{x})$ (Shapley, 1953). After that, the average contribution of all possible coalitions S excluding m can be used to calculate the Shapely value of x^m :

$$\psi_m = \sum_{S \subseteq (X \setminus \{m\})} \frac{\|S\|! (M - \|S\| - 1)!}{M!} [v(S \cup \{m\}) - v(S)]. \tag{3}$$

We use the algorithms created by Štrumbelj and Kononenko (2014) to derive the above Equation 3; Lundberg and Lee (2017) and Lundberg et al. (2018) for large M . This study uses SHAP to examine how one variable affects output based on another's value. By accurately identifying each variable's influences separately, these interaction effects can be estimated. The capability of estimating the interaction effects constitutes a featured advantage of employing Shapley values. The interaction effect for $m \neq n$ is:

$$\psi_{m,n} = \sum_{S \subseteq (X \setminus \{m,n\})} \frac{\|S\|! (M - \|S\| - 2)!}{2(M - 1)!} \delta_{m,n}[S], \tag{4}$$

Where the last item in Equation 4 is defined as $\delta_{m,n}[S] = v(S \cup \{m, n\}) - v(S \cup \{m\}) - v(S \cup \{n\}) + v(S)$.

3. DATA

The data here is sourced from the Panel Study of Income Dynamics (PSID) frequently studied by the literature on intergenerational mobility through a variety of channels such as income, education, and other factors. Following the processing procedure described in Callaway and Huang (2020) the base sample consists of 3,168 child-parent pairs after dropping observations with an extraordinary and unreasonably high level of educational attainment.

Our target is to explore the potentially-omitted relations between a child's educational attainment as the dependent variable and a list of traditional independent variables using statistical learning tools. The child's gender and birth year, parents' income, gender of the family head, race, educational attainment, and veteran status are all independent variables used in our analysis. In particular, the income used in this paper is the total family income, which includes the income of both the father and mother (Bloome, 2015; Chadwick & Solon, 2002; Mayer & Lopoo, 2005). Here, the income that is used is the permanent income, which is the average of several years' worth of income. In the literature on intergenerational mobility, constructing measures of permanent income is a major data issue (Mazumder, 2005; Solon, 1992; Zimmerman, 1992). The child's adult permanent income is determined by averaging at least three family incomes, with the condition that they must be at least 25 years old and the head or spouse of a household. When a child is under the age of 16, the average of at least three family incomes is used to determine the parents' family income. Annual family incomes of less than \$100 are omitted from these incomes before they are calculated. Using the Consumer Price Index for All Urban Consumers Research Series (CPI-U-RS) provided by the Bureau of Labor Statistics,¹ all family incomes for all years are converted into 2010 dollars. The sample is made up of people who were at least one in 1987 and were at least twenty-five in 2011, so these people are included. Additionally, to ensure that these individuals are sons or daughters at the very beginning of the survey, they must have been less than sixteen years old in 1970.

Table 1. Summary statistics.

Variables	Min.	Mean	Median	Max.	Std. dev.
Child's education	7	14.22	14	17	2.08
Parents' income	8.84	10.96	10.98	12.98	0.49
Head's education	1	2.88	3	5	1.21
Head's sex	0	0.92	1	1	0.27
Head's veteran status	0	0.39	0	1	0.49
Child's sex	0	0.5	0	1	0.5
Head's race	1	1.13	1	3	0.41
Child's year of birth	1954	1970.32	1970.5	1987	9.98

For the last step, we exclude the Survey of Economic Opportunity data component from our PSID sample; In the field of research on intergenerational mobility, this is a common practice. The most difficult part of getting the family head's variables is figuring out who the family head is. It is because the identity of the head can change over time. Throughout the whole process of their child's childhood, parents may divorce, remarry, or die. Between the time a child is born and the time the child reaches the age of sixteen, the person coded as the family head is assigned the family head characteristics as their mode of characteristics. We present in Table 1 summary statistics and Table 2 pairwise correlation coefficients between variables. As can be seen, a child's education is positively associated with the two parents' income and the head's education.

¹ The CPI-U-RS is calculated from 1978 to the present, and it incorporates most of the methodological improvements made to the Consumer Price Index over that time span.

Table 2. Correlations between variables.

Variables	Child's education	Parents' income	Head's education	Head's sex	Head's veteran status	Child's sex	Head's race	Child's year of birth
Child's education	1							
Parents' income	0.4	1						
Head's education	0.45	0.49	1					
Head's sex	0.1	0.36	0.11	1				
Head's veteran status	0.01	0.19	-0.03	0.22	1			
Child's sex	-0.06	0.02	0.01	0.02	0	1		
Head's race	-0.12	-0.25	-0.16	-0.16	-0.1	-0.03	1	
Child's year of birth	0.11	-0.05	0.31	0.04	-0.25	-0.03	-0.01	1

4. RESULTS

To get an overview of which independent variables are the key determinants of child's educational attainment for our machine learning model, we plot the SHAP value for every feature of each sample observation in the top panel of Figure 1. This scatter plot tells us whether a concerned independent variable as a predictor is positively or negatively associated with the dependent target variable. All features are first sorted according to the sum of the magnitudes of SHAP value magnitudes across all sample observations and then colored based on their SHAP value magnitudes to present the distribution of the impacts that each feature has on the model output (red means high value, and blue means low value). To be more specific, higher parent educational attainment indicates higher educated children, and higher parents' income is also a strong positive predictor of a child's education.

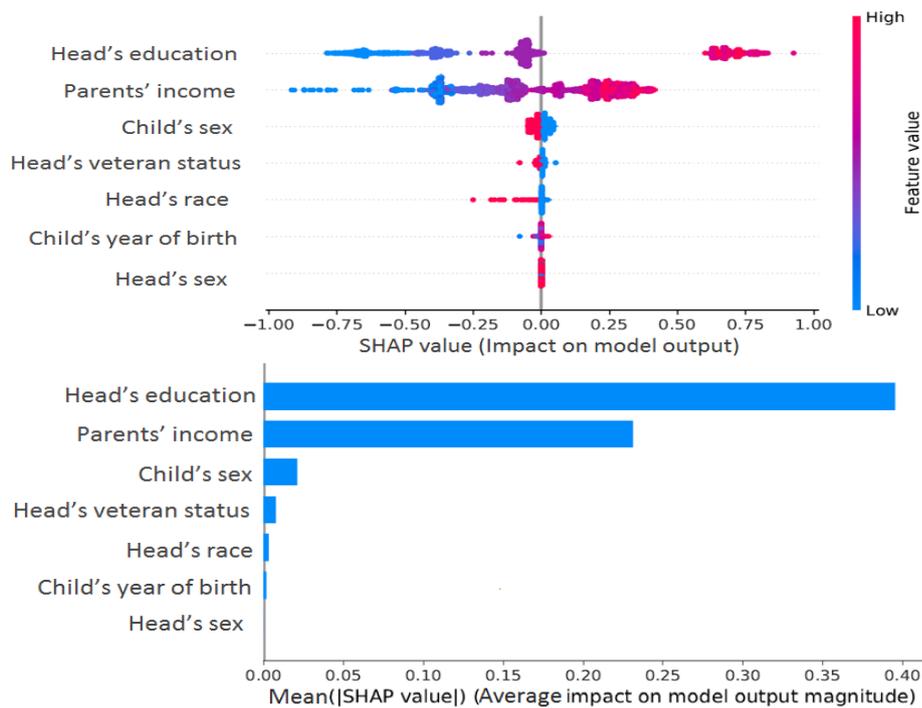


Figure 1. Aggregate and average impacts on child's education.

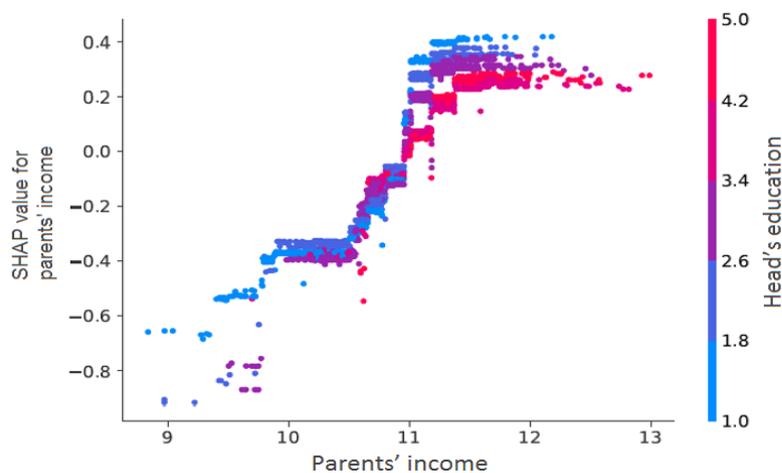
Note: This figure plots the SHAP values for all independent variables.

However, male children, children with a parent being a veteran, and non-white children are likely to experience a lower predicted education level. Children with a parent being a veteran tend to have lower education. For race, non-white child (note that the race variables are defined in a way that 1 denotes white, 2 black, and 3 others). And the rest independent variables such as the year of birth and the gender of the family head do not seem to possess any predictive power for intergenerational education mobility. In the bottom panel of Figure 1, we calculate the mean absolute

SHAP value for each feature and draw a standard bar plot that ranks independent variables by their significance as a predictor in descending order. Again, we can easily read from the bar plot that the most important feature for predicting a child's education is parent educational attainment, and the second determinant in the list is family income. Child sex and race as well as the family head's veteran status also have some predictive power for next-generation education. In sum, from a machine learning perspective, the most important determinant for a child's education is the head's educational attainment instead of the parental income—strong intergenerational education mobility.

Turning to understand how a feature interacts with the output of our model, we can plot the SHAP value of this feature against the sample value of it for all observations in the dataset. Since SHAP values represent a feature's contribution to a change in the model output (on the y -axis), this plot can represent the change in the predicted outcome caused by the change in the value of the feature (on the x -axis). Thus, the vertical dispersion evaluated at a single value of the feature represents interaction effects between this variable and a third variable. To help reveal the economic meanings of these interactions, the `dependence_plot` function in the XGBoost Python 3.7 package automatically chooses the third variable for coloring the SHAP value points under concern (this third variable appears on the right y -axis, and is selected to be the one with the highest interaction effect with the feature of interest). [Figure 2](#) includes graphical results for four features, which have high predictive power suggested previously.

The top left panel of [Figure 2](#) shows that the variable that has the highest interaction effect with parents' income is the family Head's educational attainment. Note that the other three significant features (family head education, family head race, and sex of the child) are all most interacted with the parents' income. The upward sloping trend of scattered SHAP value points demonstrates a positive relationship for intergenerational education mobility. The colors by head's education highlight the novel finding that the level of the child's education and that of their parents are positively correlated for low- and high-income families, but the correlation turns to be negative for middle-income families. Similarly, the top right panel of [Figure 2](#) shows that parental education has positive impacts on the child's education. Conditional on parent income, for families with highly educated parents, the parental income tends to be negatively correlated with the child's education and the opposite happens for families consisting of poor-educated parents. The bottom left of [Figure 2](#) says that a white child's number of years spent in school is likely to be greater than that of non-white children. However, we can also conclude that, for white and black children, their parents' income tends to be negatively correlated with the child's education; for children of other races, their parents' income turns out to be positively associated with the child's income during adulthood. Finally, in the bottom right of [Figure 2](#) we discover that female children have received more education than male children. All these interaction effect results stay consistent with the observation that the red points are located at the right side of the top panel of [Figure 1](#). In sum, regarding child's education determinants, new interaction effects among parents' income and education and other family characteristics are found by applying the proposed machine learning methodology.



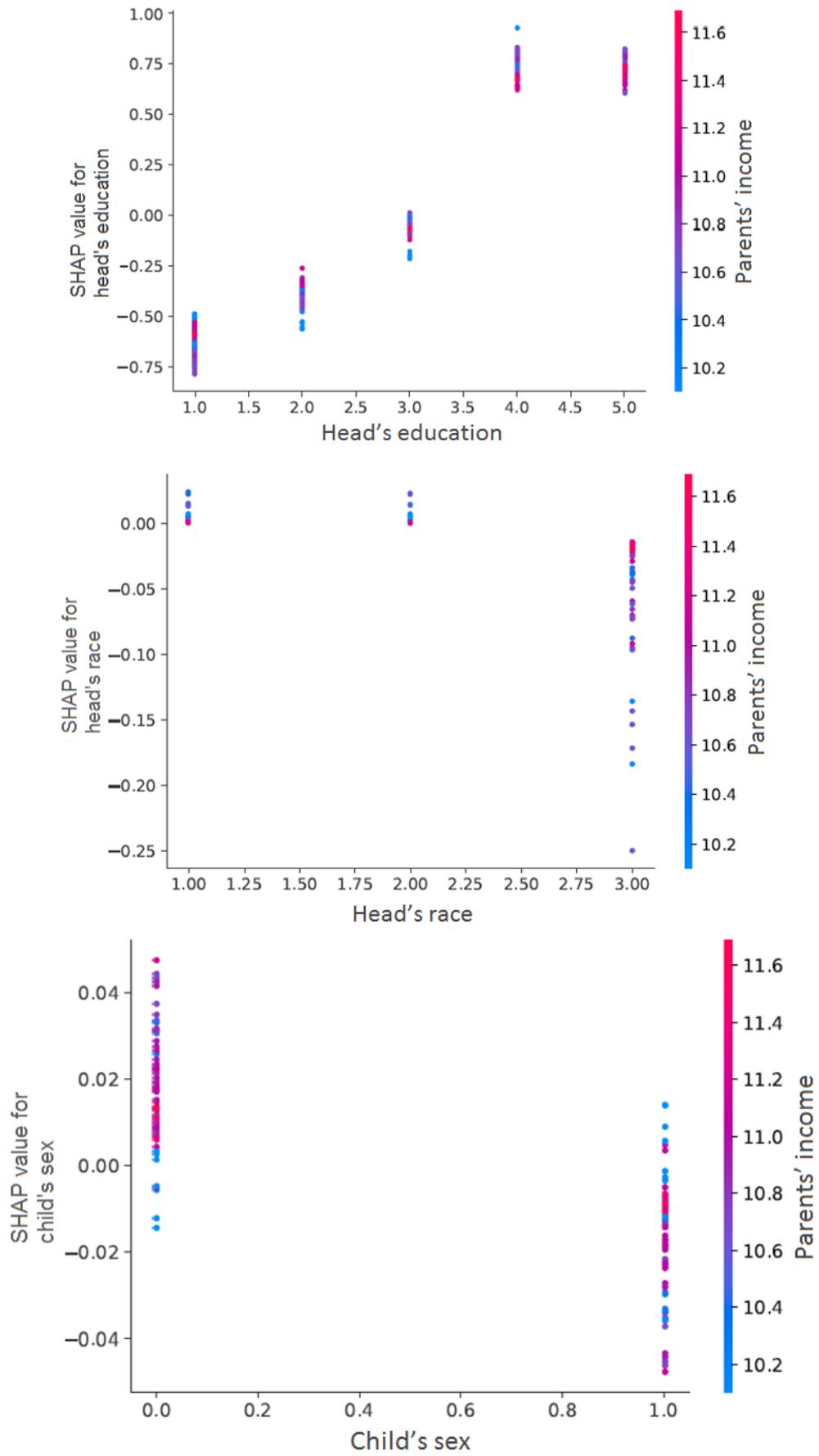


Figure 2. Interaction effects between variables.

Finally, we conduct a bunch of robustness checks using alternative machine learning methodologies including the ordinary least squares (OLS), decision trees (DT), random forests (RF), and neural network (NN). The corresponding results are included in Figure 3. Given the average impacts of factors on a child’s education estimated, it is evident that the head’s education is the most important determinant across all methods, which is consistent with the main results of intergenerational education mobility.

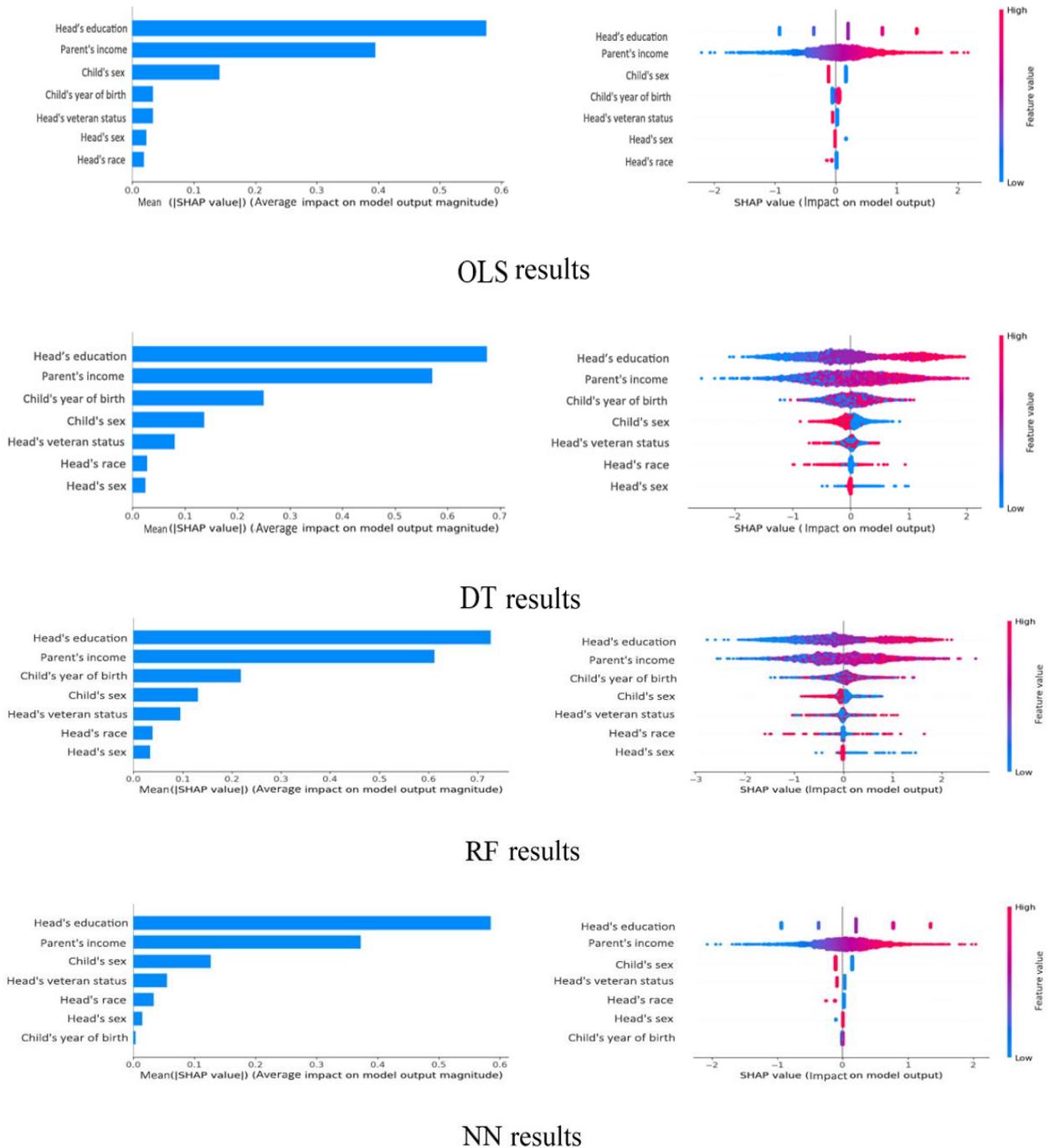


Figure 3. Average impacts child's education using various methods.

5. CONCLUSION

This article applies the machine learning methods of XGBoost and SHAP to the identification of important variables for predicting a child’s educational attainment in a frequently used dataset from the relevant intergenerational mobility literature. We find the most important predictor to be family head’s educational attainment, which is followed by the second most important predictor of parents’ income. While these results are standard, this

paper contributes to the literature by finding that the impacts of head's education on a child's education not only vary but also reverse sign with the family income level. For example, in the sample of low- and high-income families, head's educational attainment will increase the level of education of their children. But concerning middle-income families, head's education tends to be negatively correlated with the child's education. Besides, the relations between parents' income and child's education also differ according to different levels of head's education and races. All the above results have pointed directions for academics to focus on the non-linear effects of intergenerational mobility in terms of income and education. These results also imply that, if the ultimate goal is to facilitate child's education, policymakers should consider encouraging parents to attain a higher educational level rather than putting more effort in redistributing wealth to poor families.

Funding: This study received no specific financial support.

Competing Interests: The authors declare that they have no competing interests.

Authors' Contributions: All authors contributed equally to the conception and design of the study.

REFERENCES

- Aydemir, A. B., & Yazici, H. (2019). Intergenerational education mobility and the level of development. *European Economic Review*, 116(C), 160-185. <https://doi.org/10.1016/j.euroecorev.2019.04.003>
- Azomahou, T. T., & Yitbarek, E. (2021). Intergenerational mobility in education: Is Africa different? *Contemporary Economic Policy*, 39(3), 503-523. <https://doi.org/10.1111/coep.12495>
- Bloome, D. (2015). Income inequality and intergenerational income mobility in the United States. *Social Forces*, 93(3), 1047-1080. <https://doi.org/10.1093/sf/sou092>
- Callaway, B., & Huang, W. (2020). Distributional effects of a continuous treatment with an application on intergenerational mobility. *Oxford Bulletin of Economics and Statistics*, 82(4), 808-842. <https://doi.org/10.1111/obes.12355>
- Chadwick, L., & Solon, G. (2002). Intergenerational income mobility among daughters. *American Economic Review*, 92(1), 335-344. <https://doi.org/10.1257/000282802760015766>
- Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Engzell, P., & Tropic, F. C. (2019). Heritability of education rises with intergenerational mobility. *Proceedings of the National Academy of Sciences*, 116(51), 25386-25388. <https://doi.org/10.1073/pnas.1912998116>
- Ferrare, J. J. (2016). Intergenerational education mobility trends by race and gender in the United States. *AERA Open*, 2(4), 1-17. <https://doi.org/10.1177/2332858416677534>
- Fletcher, J., & Han, J. (2018). *Intergenerational mobility in education: Variation in geography and time*. Retrieved from NBER Working Paper No. 25324:
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>
- Jungert, T., Levine, S., & Koestner, R. (2020). Examining how parent and teacher enthusiasm influences motivation and achievement in STEM. *The Journal of Educational Research*, 113(4), 275-282. <https://doi.org/10.1080/00220671.2020.1806015>
- Leone, T. (2019). Intergenerational mobility in education: Estimates of the worldwide variation. *Journal of Economic Development*, 44(4), 1-42.
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56-67. <https://doi.org/10.1038/s42256-019-0138-9>
- Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *Methods*, 5(13), 1-9.

- Lundberg, S. M., & Lee, S. I. (2017). *A unified approach to interpreting model predictions*. Paper presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems.
- Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., & Adams, T. (2018). Explainable machine-learning predictions for 330 the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10), 749-760. <https://doi.org/10.1038/s41551-018-0304-0>
- Mayer, S. E., & Lopoo, L. M. (2005). Has the intergenerational transmission of economic status changed? *Journal of Human Resources*, 40(1), 169-185. <https://doi.org/10.3368/jhr.xl.1.169>
- Mazumder, B. (2005). Fortunate sons: New estimates of intergenerational mobility in the United States using social security earnings data. *Review of Economics and Statistics*, 87(2), 235-255. <https://doi.org/10.1162/0034653053970249>
- Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. K. (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention*, 136, 105405. <https://doi.org/10.1016/j.aap.2019.105405>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Why should I trust you? Explaining the predictions of any classifier*. Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Russino, A. (2018). Financial development and intergenerational education mobility. *Review of Development Finance*, 8(1), 25-37. <https://doi.org/10.1016/j.rdf.2018.05.006>
- Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28), 307-317.
- Solon, G. (1992). Intergenerational income mobility in the United States. *American Economic Review*, 82(3), 393-408.
- Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3), 647-665. <https://doi.org/10.1007/s10115-013-0679-x>
- Turcotte, M. (2011). Intergenerational education mobility: University completion in relation to parents' education level. *Canadian Social Trends*, 92, 37-43.
- Zimmerman, D. (1992). Regression toward mediocrity in economic stature. *American Economic Review*, 82(3), 409-429.

Views and opinions expressed in this article are the views and opinions of the author(s), World Journal of Vocational Education and Training shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.